STATISTICS FOR ECONOMISTS

by

R. G. D. ALLEN,

O.B.E., M.A., D.Sc.

PROFESSOR OF STATISTICS
IN THE UNIVERSITY OF LONDON

HUTCHINSON'S UNIVERSITY LIBRARY

11 Stratford Place, London, W.1

New York *Melbourne* *Sydney* *Cape Town*

THIS VOLUME IS NUMBER 13 IN
HUTCHINSON'S UNIVERSITY LIBRARY

First Published . . *Feb.*, 1949
Reprinted . . *Feb.*, 1949

Printed in Great Britain by
William Brendon and Son, Ltd.
The Mayflower Press (late of Plymouth)
at Bushey Mill Lane
Watford, Herts

CONTENTS

*These sections involve more advanced ideas and they can be left for a second reading of the whole text.

CHAPTER I

THE RAW MATERIAL

1.1 *Introduction.* We can do no better than to turn to an authoritative dictionary for a concise definition of statistics:

> *Statistics.* Noun, plural. Numerical facts systematically collected, as statistics of population, crime; (treated as singular) science of collecting, classifying and using statistics. (*Concise Oxford Dictionary*, 1929 edition.)

The term statistics can then be used in the plural to denote collections of numerical data and in the singular to refer to the technique of statistical analysis. Here we shall attempt to keep the two things separate by using statistics for the data and statistical method or analysis for the technique.

Statistical method is essentially a branch of mathematics, using the processes of reasoning which make up scientific method in general. Many, perhaps most, expert statisticians are mathematicians by training or adoption. There are other experts who make up for a lack of mathematical equipment by an extensive and intensive knowledge of particular fields of application, e.g., as economists and sociologists as much as statisticians. They may not have knowledge of more than elementary statistical methods, but they know where to seek the statistical data they need and how to use the data.

A knowledge of statistical methods is not only essential for those who present statistical arguments; it is also needed by those on the receiving end. The economist, administrator or politician requires statistics to support his arguments and to illuminate the problems he handles. The general citizen, if he is to be enlightened, must refer to many books and articles which employ statistics. But these must do more than read the text and look at the statistical tables and charts; they must be able to understand what the statistics mean, appreciate their limitations and criticize their use. They must follow the discriminating middle course between the extremes of the cynic who thinks statistics can prove nothing and the uncritical

believer in the veracity of every figure. They should be able to turn their hands to a little statistical manipulation on their own, if only to check the statistical analysis presented to them. It is primarily to these that the present account is directed.

The ordinary user of economic and social statistics gets his material ready made from an official or private publication; or at most, he gets it in the rough and adapts to his own needs. Very seldom indeed does he need to go out and collect his own data at source. It is like buying a suit of clothes "off the peg" or through a tailor, not by going back to the sheep or even by weaving wool yarn into cloth. But, unlike the purchaser of the suit, the user of statistics must know how his material has been worked up. If he is not to fall into grievous error, he must discover how his data have been collected, classified, put up into tables and presented to him. We must, therefore, spend a little time on these topics.

1.2 *Statistical Inquiries.* The collection of statistical data is largely and increasingly the function of government. The smooth running of official machinery depends on an adequate supply of statistics, and the wider the operations of government, the greater the volume of statistical data needed. There was a time, not so many years ago, when governmental control and regulation of the national economy was more limited and large gaps in official statistics were filled by commercial organizations and independent investigators. Private statistics were first in the field in the development of statistical information on national income and outlay, on wages, earnings and family expenditure, on housing and social conditions of the people and other vital topics. The pioneering work of Bowley, Stamp and Colin Clark, of Charles Booth and Seebohm Rowntree, and of many others has not only provided us with invaluable data, but has been of great benefit to official statisticians when government departments took over responsibility for the regular collection of data. Even now, the needs of business and social investigators are by no means fully met from the storehouse of official statistics. Large firms still collect their own data and so do commercial organizations such as those engaged in advertising and market research. There is still room for private research groups such as the National

Institute of Economic and Social Research, and for *ad hoc* investigating committees like that which conducted the *New Survey of London Life and Labour*.

We can, however, take official statistics as typical of the rest. A very large amount of statistical information is collected by government departments as a matter of routine and as an essential basis of administration. A good deal is also published for the general information of the public as well as of officials. Indeed, the Government recognizes increasingly its responsibility for the provision of data in adequate volume and suitable for use by outside investigators, and for the information of the general citizen. The collection and publication of such statistics as the decennial population census, the annual crop estimates and the monthly figures on employment are now an accepted and unchallenged duty of government. Other inquiries, such as the annual census of production and regular estimates of earnings of workers, are now undertaken officially, and will soon be equally accepted by industry and by the public. In addition to regular collections of data, information for policy guidance or for the illumination of particular problems is assembled *ad hoc* by official agencies or committees; the inquiry into working-class expenditure in 1937–8, the National Farm Survey of 1941–3 and the Family Census of 1946 are examples.

The design of an original inquiry, the setting up and, later, the modification of machinery for regular collection of data are operations not to be lightly undertaken. Time and money can be saved and accuracy improved by careful and detailed planning in the initial stages. A good example is provided by the way in which the census of production was resumed after the War of 1939–45. The whole question of what information should be obtained was first referred to a committee of officials and outside experts appointed by the President of the Board of Trade in 1945. An Advisory Committee was then constituted to co-operate with the statisticians of the Board of Trade to implement the recommendations of the first committee. Some months were taken in framing instructions, in drawing up specifications and in designing forms, always in consultation with representatives of the various industries concerned. A partial census was taken in 1947, relating to business in the

year 1946 in certain selected industries. Finally, in the light of this experience, modifications are being made in the procedure for the first full census to be taken in 1949 in respect of the year 1948.

In short, as complete a plan as possible should be drawn up before the actual collection of information begins, specifying what statistics should be obtained, from whom and by what methods. There should also be full and precise definitions of terms, instructions to investigators and respondents and some indication of the lines on which the analysis of results is to proceed. Though the plan should be complete, it should not be completely rigid. Adjustments are inevitable as collection and analysis of data proceed, and the plan must be adapted to the conditions actually found.

The expenditure of money and of time, which is often the same thing as money, is always a factor, and usually a dominant factor in the conduct of an inquiry. Like a problem in engineering, a statistical investigation must be designed to do a stated job with minimum expense and in the shortest time, or conversely to get the best results from a given sum in a given time. The complete job includes both the collection of data and the preparation of results for use and publication. It may be a matter of £100 and a few weeks—or of thousands of pounds and several years. But the same principles govern the design. Clearly, at this stage the statistician must be an administrator and must co-operate with many others.

A fundamental question which turns largely on the matter of expense is whether a *census* with a complete enumeration of the whole field is to be adopted or a partial survey of part of the field involving a selection or *sample* of the whole set of items. A census may lead to greater accuracy and more refinement in analysis, but it can be a very expensive and lengthy operation. A sample designed and taken with skill can produce results which may be sufficiently accurate for the objects in view, and it can save much time and money. With the development of statistical techniques of sampling, and with greater knowledge of what samples can achieve, the reluctance of administrators and investigators to use samples is being overcome and the sample is increasingly employed in official as well as in private inquiries. There will, however, always be

scope for the census method in cases where the field surveyed is not extensive or where, as with the census of population, need for a comprehensive coverage on certain details is imperative. In the latter case, a combination of the census and sample methods can be adopted and has much to recommend it. So, frequent sample surveys can be used to fill the gap between censuses taken at intervals. Or, certain simple questions can be asked of every one, while other and more complicated questions can be put only to a proportion (say 5 per cent) of all respondents in a census, as in the U.S. Census of Population of 1940.

1.3 *Forms and their Design.* Having decided on a census or on one of the many possible forms of sampling, the statistician administrator has still to settle a whole host of questions. In economic and social inquiries, information is almost always collected by getting someone to fill up a form or questionnaire. Every citizen must inevitably participate in filling up many forms in the course of his life-time, for births, marriages and deaths are so recorded as are the composition of households in the decennial census and registration for the exercise of the basic right to vote. If he makes his living as a wage or salary earner he fills up forms for his social insurance and to ensure that he pays his income tax. If he operates a business he has scores of forms to handle every year. All these are designed directly, or employed indirectly, to provide statistical data. One matter to be decided is whether the forms should be filled up by an enumerator or investigator who collects data by asking questions and noting down answers, or whether they should be left with the respondent to complete on his own. Public opinion polls are instances of the former and the income tax schedule of the latter. Sometimes, as in the population census, a combination of the two is adopted, some part of the form being filled up by the respondent and others by an enumerator.

Whichever method is used, the design of the form and of the accompanying instructions to enumerators and respondents is a matter to be decided with great care. Each question must be clearly phrased and capable of unambiguous answer. Instructions must cover all possibilities, even remote ones. One of

the difficulties is that the way of "popping the question" can influence the answer as all those who have tried to poll public opinion will acknowledge. The difficulties arise even in a relatively simple matter such as the determination of age. Suppose that an inquiry needs only to separate those under eighteen from those aged eighteen and over so that a direct question such as "Are you under eighteen?" would seem sufficient. Such a question may easily produce biassed answers if respondents think there is some advantage in answering one way or the other. It is safer to put the question in the form "State your age at last birthday," or "Give date of birth." This leaves less chance for misrepresentation or faulty arithmetic on the part of the respondent. Another device is to insert a question primarily to produce answers which can be used to check the answers to other questions, as age may check the stated relationship of one member of the household to another. It is for these reasons that forms often appear to include unnecessary and irrelevant questions. Most publications of good statistical inquiries reproduce the layout of the form used; all should do so. This is not a waste of paper for the form should be studied critically by all who use the data.

1.4 *Reliability of Data.* All these things and many others are to be taken into account in assessing the reliability of basic material as collected. Such an assessment must be made before the data are analysed and conclusions based on them. The job of the statistician is to get material of the greatest possible accuracy and then to make the best use of it. He should not, however, discard imperfect data if nothing better is available. Not even the most subtle and skilful analysis can overcome completely the unreliability of basic data, but the best can always be made of a bad job. Some material, so rough as to be insufficient for fine analysis, may still support particular conclusions. The skilled statistician knows where he can proceed and where he must stop.

Speaking broadly, we can say that accurate data is obtained only when those supplying information appreciate the need for it. It is better if they have something to gain from accurate answers, or if they are subject to enforceable penalties for failure to respond correctly. Information is less reliable if

respondents are not interested, or if they are not approached skilfully, and less reliable still if they are definitely hostile to the objects of the inquiry. For example, a large industrial firm can handle efficiently the forms of a census of production or the returns required on employment; responsible business men know that these are necessary and they have been consulted in advance through their trade associations. On the other hand, a small shopkeeper is not so well informed; he may say that his job is to sell his goods efficiently and to his greatest profit, not to fill up forms at the pleasure of bureaucrats. Again, the census of population in a country like ours is a well-established and recognized procedure, backed by legislative powers for enforcing penalties on those failing to respond. Even so, there are some minor uncertainties, for example in the records of women's ages, and occasional difficulties of major importance, as in 1911 when the organized suffragettes refused to co-operate in the census. To appreciate the real achievements of our census, we can ask how we would set about taking a census in a colonial area where some members of the population may be difficult to locate, where they are ignorant of the real purpose of the census and where they may not even know many of the answers required of them. The results of such a census may be wildly different, for example, if it is thought that the object is to enumerate people for issue of ration books on the one hand, or for assessment to poll-tax on the other.

All statistical material involves errors in its collection, varying from minor slips in response to actual misrepresentation. In addition, as we shall see later, the sampling method introduces errors of a different and more controllable type, those arising from the fact that one selection of items has been made rather than another. We can note here that the two types of error are not necessarily additive. A census requires many enumerators and more respondents, and it may be spread out over a considerable time with the result that errors of the first kind may be quite large and frequent even if the questions are limited to a few very simple ones. In a sample, it may be possible to reduce such errors by intensive training of a few enumerators and by more patient methods of approach to respondents. At the same time, the sample

inquiry may introduce a wider range of more complicated questions. The census does not always have the edge over sample in the matter of accuracy.

There are many other points to be considered in judging the methods of collection of data. The actual procedure adopted in field work or in distributing forms needs to be examined. If the forms are sent out and returned by post, what is done to see that every one entitled to a form actually receives one? What steps are taken to "follow up" all those who do not reply? If the inquiry is by interview, how long does the enumerator spend over each? What instructions is he given about re-visiting those absent at the first call? Is the work of different enumerators checked, for example, by replication of interviews whereby each respondent is interviewed by more than one enumerator? Again, the particular circumstances under which an inquiry is made, even if apparently irrelevant, may be highly important in assessing the outcome. For example, the results of the inquiry into earnings conducted by the Ministry of Labour in January, 1945, were affected by the fact that the weather at the time was particularly severe. The population census of 1921 was postponed from the usual date in April until June of that year, with consequent differences in the geographical distribution of the population. The numbers recorded in Blackpool or Brighton in 1921 cannot be compared directly with those recorded in 1911. Such examples can be multiplied indefinitely, but enough has been said to make the point that the assessment of any statistical data is a matter which is neither easy nor learned by any other process than that of hard experience.

1.5 *Summarization of Raw Data.* The result of the collection of statistical data is a mass of forms or returns. The next step, clearly, is to examine each return, to edit and to check it for internal consistency. Omissions and errors will inevitably be found and some of the returns may need to be returned to the originator for confirmation, alteration or amplification. At this stage, too, anything radically wrong with the method of collection will be detected (e.g., fradulent returns from particular respondents or enumerators) and part of the work may need to be re-done. After editing has been completed, the

statistician still has a mass of forms on his hands; he must proceed to bring order out of chaos and he must begin to see the wood for the trees. What is needed is a summarization of the raw material, a long process which ends with the presentation of a set of statistical tables with an accompanying text and commentary. Appendix I gives a set of fourteen tables of various types for illustrative use here and in the later developments of statistical methods.

Summarization of statistical data into tabular form is an art rather than a routine following a set of formal rules. Tabulation inevitably implies a loss of detail. The original data are far too voluminous to be appreciated and understood; the significant details are mixed up with much that is irrelevant. The art of tabulation lies in the sacrifice of detail which is less significant for the purposes in hand so that what is really important can be emphasized. Tabulation implies classification, the grouping of items into classes according to various characteristics. And classification depends on clear and precise definitions.

1.6 *Definitions.* The main definitions will have been specified in planning the inquiry; they will be amplified and modified in the process of classification and tabulation. As we shall always emphasize, statistical methods are basically practical and common-sensible, but rather more precise and systematic than is usual in everyday life. Our definitions, then, must be based squarely on common usage, but it is equally important that we see that they are precise and unambiguous, that they are uniform and leave no gaps. It follows that, in statistical definitions, we must sometimes be arbitrary and we may have to split hairs. A term can be employed in ordinary use without bothering unduly about little inconsistencies or loose interpretations. This won't pass in statistical work where every item must go into one group or another and where nothing must be left on the border-line. In fact, it is the treatment of border-line cases which makes statistical definitions arbitrary and sometimes laughably so. Though definitions may approach the ludicrous, we must accept this as the price which may have to be paid for precision.

For example, a standard list of industries grouped into

appropriate classes is essential in statistical presentation. Motor garages must be put in one of the classes. But these useful businesses perform several functions; they repair and rebuild cars, they store vehicles, they sell petrol and oil and they provide a variety of services from wiping the windshield to polishing the body and lubricating the transmission. They may, therefore, be classed with the motor industry, along with Austin or Nuffield, they may be included with road transport, they may go in as a service trade or they may be put in other groups. In fact, at different times and for various purposes, they have been put in several groups. But it is clearly preferable to be rather arbitrary and to place them definitely in one group for all purposes. Otherwise we cannot be certain we are comparing like with like when we relate, for example, employment in an industry with the output of that industry. A classification of persons by occupation raises the same problems and involves some arbitrary separations. Looking at the classification of the 1931 Census of Population, we may agree that tipsters and tic-tac men go logically with bookmakers, but we may wonder why bellringers are musicians and piano-tuners makers of musical instruments, or why railway detectives are police, while private detectives are engaged in personal service.

The tables of Appendix I provide a variety of illustrations of definitions, some simple and some more complex. To take one example of a complicated kind, we see that Table 2 is based on the definition of the "working population" of Great Britain and on the division between those attached to some industry and those not in employment. There are common-sense distinctions to serve as guides; schoolchildren, full-time students, housewives not otherwise employed, and retired persons are not in the working population, while employers, self-employed, wage-earners and salaried workers are. This is not quite good enough, however, since there are many doubtful cases. How do we treat the shopkeeper's wife who sometimes minds the shop or the farmer's daughter who helps on the farm? What of the married woman who "chars" in the afternoon or the boy who delivers the papers before school? The Ministry of Labour answers such questions in defining the data of Table 2. First, only males aged fourteen to sixty-four

and females aged fourteen to fifty-nine are included for administrative reasons, since sixty-five and sixty are the pensionable ages of men and women respectively. Secondly, all employers, all self-employed persons and all in *paid* employment are included apart from private domestic servants (again for administrative reasons). This rules out the shopkeeper's wife and farmer's daughter if they are not paid. Thirdly, those absent from work at a particular date because they are sick, on holiday or just taking a day off are counted as employed if maintained on their employer's books. Those classed as not in employment are insured workers registered at Employment Exchanges and ex-Service personnel who have not yet taken up employment. Hence, a worker who has been sick so long that he is neither on the books of his old employer nor registered at an Employment Exchange is excluded from the working population. Finally, a woman in part-time paid employment, defined as thirty hours a week or less, is counted as half a full-time worker. This brings in the office "char" but not the household help who is excluded as a private domestic servant. There are very few men in part-time paid employment; those on the civil staffs of government departments are counted as halves, but otherwise no distinction is drawn between part-time and full-time workers.

1.7 *Classification.* With definitions clearly laid down, we can proceed to classification and tabulation in the form and detail suited to the purposes of the presentation. Ideally, a class or group should be homogeneous; that is, it should include all items, and only those items, with a definite characteristic. This is rarely possible in practice and most statistical groups are somewhat "mixed bags" of more or less heterogeneous items. The object, however, is to make them as homogeneous as possible with regard to significant characteristics while accepting heterogeneity in less important respects.

The more usual types of classification and tabulation can be illustrated by reference to the tables of Appendix I. We can distinguish, first, classification by a qualitative or non-measurable character, or *attribute*, from cases where a measurable character or *variable* is involved.

Tables 1, 2 and 3 are quite simple examples of classification

by attributes (the rows of the table) repeated at different dates. The net national income of the U.K., £4,671 millions in 1938, is classified in Table 1 both by the sources of income and by the various ways in which income is expended. In distinguishing the ways in which income is earned, for example, we have rent, interest and profits, salaries, wages and the income, in cash and in kind, of serving members of the Forces. Transfer incomes such as unemployment and sickness benefits, and also interest on the national debt, do not appear since net national income is defined to exclude them. The definitions of the different categories of income are somewhat complex and obtainable only by reference to the publications from which the data are derived. In particular, the distinction between wages and salaries approximates but does not correspond exactly to that which may be made in employment data between manual and non-manual workers. When information is given at different dates, as in Table 1, it is not worth while attempting to improve the classification (e.g., the distinction between wages and salaries) unless the revisions can be carried back to all dates. It is more important to have the *same* classification at all dates, even if it is not exactly what is required. Table 2 is similar to Table 1; the total classified is the working population and the classification is by broad industrial groups.

Table 3 shows the quantities and values of U.K. exports in one of the categories—beverages and cocoa preparations—distinguished in the official trade returns. The classification is according to commodity and three classes of drink and two of cocoa preparations are shown. The latter are described officially as cocoa preparations (not containing spirits)—containing sugar and not containing sugar respectively. Less exact but more recognizable labels for the two categories would be chocolate and cocoa. A miscellaneous group entitled "all other items" appears in this as in many other tables; here it is not given by quantity and an entry is made only in the value columns. Such a miscellaneous group should be made as small as possible relative to the total and, in this instance, it is a little over 5 per cent of the total in 1946 and smaller in other years. The lay-out of Table 3 is also designed to facilitate further calculations on the basic figures. Table 4, even more than Table 3, is essentially a work-sheet and the main interest is in the

index number which is the final product of the computations.

1.8 *Time Series and Frequency Distributions.* Tables 5, 6 and 7 of Appendix I provide examples of time series in which a variable is given at monthly, yearly, or other regular intervals. Three series are shown in Table 5, yearly, over a long period, and one of them is derived from the other two. The four series of Table 6 are also given yearly. Table 7 consists of a single time series of monthly figures; it differs only in that the figures are arranged in rows and columns, the whole series being read down successive columns. Table 8 is similar to a time series, but the variables are now given, not at successive points of time, but for different places or areas.

The remaining tables of Appendix I provide examples of frequency distributions in which a set of items (e.g., families) is classified according to the values assumed of one or more variable characteristics (e.g., income, food expenditure). Table 9 shows how two alternative distributions can be formed from the same basic data. The items distributed are the price relatives, forty-five in number, used in the *Statist* index number of wholesale prices; these are shown first in their original order, then re-written in ascending order of magnitude and finally grouped in alternative frequency distributions. The first of these distributions is to be read: two items in the range of price relatives from 0 to 24 inclusive; three items in the range from 25 to 49 inclusive; eleven items in the range from 50 to 74 inclusive; and so on.

Table 10 relates to a group of families for each of which income and expenditure on food are given. The families are distributed, first according to income alone, in classes of specified income ranges, and then in a double (or square) frequency distribution, the families being distributed both as to income and as to food expenditure. One "border" of the double table, i.e. the bottom row, is a condensed version of the simple income distribution; the other "border"' i.e., the right-hand column, is a similar simple distribution of families according to food expenditure.

It is essential in a frequency distribution to keep the items which are distributed clearly distinguished from the variable

of the classification; the latter gives the ranges of the classes and the former provide the entries in the table. In Table 10B, for example, the items distributed are ninety families and the entries in the table are numbers of families; the variable character from which the classes are formed is family income. The table is the distribution of families according to income.

The definition of the classes selected for the variable needs to be precise to avoid ambiguity with border-line cases and, for this, a clear notation is required. There is little difficulty when the variable takes only a limited number of discrete values, as the number of wage-earners in the coal mines of Table 13. Here the ranges of the classes should be given inclusive as follows:

1–19; 20–49; 50–99; . . .
not as: 1–20; 20–50; 50–100; . . .

The first notation makes it clear that 20 is in the second class and not in the first; the second notation leaves this point unsettled.

There is more difficulty when the variable is continuous like age, or approximately so like income. We need to state first how the variable is measured, since even for a variable which is theoretically perfectly continuous, it is estimated in practice only to a certain fineness. So, age may be given to a month, income to a shilling or a price relative to one decimal place. The definition of classes of the variable should then take this into account and should adopt a notation which leaves no ambiguity about the range of each class. Two examples will make the main points.

In Table 9 the price relatives are each rounded off very severely, in this instance, to the *nearest* whole number. The figure 53, for example, means that the price relative lies anywhere between 52·5 and 53·5. If, by chance, the relative is exactly at the halfway point, say 52·5, then by convention we may round it upwards to 53. So, the range given is from 52·5 to 53·5, including the first but not the second. Having stated the rounding to the nearest whole number, we specify the classes of the distribution in Table 9 as:

0 and under 24·5; 24·5 and under 49·5;
49·5 and under 74·5; . . .

which can be written more shortly but without ambiguity:

0–24; 25–49; 50–74; . . .

but not as:

0–25; 25–50; 50–75; . . .

Age in Table 11 is a continuous variable which may be taken as given (e.g., by date of birth) to the day. In defining the class ranges, we must indicate clearly whether we put an exact year (those with birthdays on the day of the count) at one end or the other end of the range; it cannot be both. In this instance, the exact ages are put at the lower end and the classes are:

Under 5; 5 and under 15; 15 and under 25; . . .

For a shorter notation, we can write either

0– ; 5– ; 15– ; . . .

or

0–4; 5–14; 15–24; . . .

The former is preferable and adopted here. The latter, as used by the Registrar-General, may be confused with the case when age is rounded to the nearest year. It can be used safely, however, if it is stated and understood that age is given at *last* birthday. We can notice that we could include the exact year of age at the upper end of the range, by taking age at *next* birthday, and appropriate classes would then be:

Not over 5; Over 5 and not over 15;
Over 15 and not over 25; . . .

This can be written either

–5; –15; –25; . . .

or

1–5; 6–15; 16–24; . . .

In neither case should we use the notation

0–5; 5–15; 15–25; . . .

which is ambiguous as to the allotment of the exact years 5, 15, 25 and so on.

1.9 *Tabulation.* It is only by experience that skill is acquired in the framing of tables. It is partly a matter of design, to get a neat and concise layout which is both cheap to print and easy on the eye. It is partly a question of making sure that no essential information is omitted so as to leave the meaning

of the table uncertain. No general rules can be given but the following points are to be borne in mind.

The heading of the table should be short and confined to a brief description of what the table is about with a note of such things as date or place which may be common to all entries. Similarly, the heading of a column (or row) should indicate concisely what the entries are and, in particular, what distinguishes one column (or row) from another. Somewhere in these headings there must appear the dates and the places to which the contents of the table relate. Further, there must be a specification of the nature of the entries and of what units are used, e.g., £millions in Table 1 or population in thousands in Table 2. Equally, in a frequency distribution, there must be a definition of the variable classified into ranges, e.g., age in years in Table 11.

In designing a table, it is advisable to draft a blank form first, to see how it looks and, by inserting notional entries, to find out if it works. We can err on the side of over-elaboration, or of impossible condensation. As an extreme example of the latter, suppose we have records of incomes and rents for two groups of families, one in Manchester and the other in Liverpool. We shall soon find it impossible to fill up a table as short as:

	Income	*Rent*
Manchester		
Liverpool		

We cannot enter numbers of families in this table, nor £'s of income and rent. We could, however, extend the table by showing ranges of income and rent and then entering numbers of families. Or we could push on a stage further with the statistical analysis and write average income and average rent at the top of columns containing £ figures.

Since the headings must be concise, it is almost inevitable that a good deal of important information will be left over to be given, at greater leisure and in finer print, in footnotes. These can conveniently cover such matters as detailed specifications, definitions of terms and particular qualifications of figures. Finally, the source of the data utilized must be indicated clearly in the table, even if they are taken from a well-known publication. This is not an idle point for it is

essential that the user of a table should be given the chance to go to the original source and to check or amplify the data for himself. Moreover, a reference to the source makes it possible to omit much of the detail of definition and classification. The reader is left, as in Table 1, to obtain full and exact particulars from the source quoted.

The process of getting the raw data into a table designed to take them is straightforward but arduous. When the number of items handled is not large, as in Table 10, where there are only ninety families, tabulation by hand is simple and rapid. Some effective checks on arithmetic should be applied, e.g., by seeing that the entries always add to the same total and that columns and rows add down and across. The safest procedure is to get the tabulation performed by two independent computors.

If a much larger number of items is to be tabled, tens or hundreds of thousands rather than a few score, the work of hand tabulation soon becomes prohibitively long and uncertain. Mechanical aids to tabulation and computation have been designed; they are constantly improved and increasingly used in statistical inquiries. For addition, multiplication and division, there are machines operated by hand or by electricity. For tabulation there are larger and more complicated electrical machines. The essential basis of machine tabulation is the transcription of the original data, usually on forms filled up in manuscript of varying legibility, to specially prepared cards which summarize the information concisely case by case. The usual method is to punch holes at the appropriate points of designated columns on the card and then to pass the cards through the machine which separate them into groups or record the information on them by means of electrical contacts through the holes. Machine tabulation is described in the technical literature.[1] The recording of statistical facts has kept up with modern trends of mechanization. It is a long way from the tally sticks of the Treasury to the tabulating machines used by the Registrar-General.

[1] *See*, for example, J. P. Mandeville, "Improvements in Methods of Census and Survey Analysis," *Jour. Roy. Stat. Soc.* (1946).

CHAPTER II

SOURCES OF PUBLISHED STATISTICS

2.1 *General Sources.* This is the point where the ordinary working statistician in the economic and social field comes in. He seldom collects his own statistical material; instead he makes use of tabulations and reports prepared and published by others. The data have been collected, classified, tabled and presented to him with text and commentary. But his job is still an important one, requiring skill and knowledge. He must define precisely the object of his particular inquiry, determine what data he wants and how he needs them classified. Next he must know where to go to get the data and how to compromise between what he wants and what he can get when, as often happens, information is not available in precisely the form needed. He must examine, closely and critically, the nature and derivation of any material he takes over, with particular emphasis on its reliability and limitations for his purposes. He must then proceed to assemble, arrange and condense the data, usually given in his source in tables too detailed for his object and not set out in the most suitable way for him. Finally, he must draw his conclusions and present his results in a report or article which will contain as much, or as little, tabular statistical matter as he thinks his readers can absorb. The notion that a "literary" presentation of facts and figures is easier to read is dying out and most investigators now feel at liberty to give their readers a set of tables which can be more concise and more accurate than the clearest of prose styles. It is evident, then, that such a working statistician must possess qualities of no mean order—he must have wide knowledge of his sources, considerable skill in technical statistical analysis and an honest and unbiassed approach to his problems. Something will be said about sources in this chapter. An outline of the simpler statistical techniques follows in subsequent chapters. The honesty of the investigator must be assumed throughout.

Two small examples may be given to illustrate the difficulties in handling the sources of data. Suppose we wish to determine the proportion of the national income derived as profits of industrial and commercial undertakings. In the official source of over-all data on national income, as used in Table 1, we find a group entitled "interest and profits, including farming profits and professional earnings," more fully described in technical notes. This is too wide for our purpose and no analysis into constituents is available. We must either "make do" with this item, or look elsewhere. Again, suppose we wish to follow the course of employment amongst workers of a given category over a period of years. We will find that we cannot get completely comparable figures since the definitions of the group of workers, and the methods of recording the numbers employed and unemployed, have changed at different times. This is the case with the data of Table 6, which show a major change in 1938 making later figures non-comparable with earlier ones.

The practising statistician is always in something of a dilemma. He is anxious to get his basic data improved, e.g., a more complete coverage or a classification better suited to his purpose. On the other hand, he is interested in seeing that his data are on the same basis and so comparable over time. His best compromise is usually to get changes incorporated at quite frequent intervals, with a double calculation shown at the time of any change (first on the old and then on the new basis) so that he can "splice" the new series on the old to make them approximately continuous.

There are several general sources to which the investigator can refer for a wide range of British official and private statistics on economic and social matters. The *Statistical Abstract* gives "potted" figures on most topics and generally for a period of fifteen years. This publication was issued annually by the Board of Trade up to the War of 1939–45, the last issue being the eighty-third number covering the years 1924–38 inclusive (Cmd. 6232). In January, 1948, it was revived by the Central Statistical Office under the title *Annual Abstract of Statistics*. *The Monthly Digest of Statistics* gives a rather smaller range of data in monthly series. It was first issued in January, 1946, by the Central Statistical Office, and it is accompanied by a

valuable pamphlet on "Definitions of Items and Units" giving the primary sources and short descriptions of the monthly series. Other official periodicals containing regular statistics are the Board of Trade *Journal* (weekly) and the Ministry of Labour *Gazette* (monthly). The latter can be used in conjunction with the *Abstract of Labour Statistics* issued by the same department at intervals until it was suspended with the twenty-second number covering the years 1922–36 inclusive (Cmd. 5556). The annual White Paper on *National Income and Expenditure* (cf. the issue of 1947, covering the years 1938–46, Cmd. 7099) contains much information on the British economy as a whole. An annual series of White Papers was inaugurated with *Economic Survey for* 1947 (Cmd. 7046) to marshal broad statistical material for purposes of diagnosis and forecast of the economic situation.

The location of official reports on particular topics is not an easy matter in view of the volume of official publications. The general publications mentioned above are themselves invaluable as references to more detailed reports and tabulations. In addition, the following official works of reference are designed to assist the investigator in his search: the annual *Guide to Current Official Statistics* (suspended with the seventeenth volume, 1938), the *Consolidated List of Government Publications* published annually by H.M.S.O., and the lists of *Parliamentary Papers* issued each session as a House of Commons Paper.

Amongst private publications containing regular British statistics, official and other, there are the quarterly *Bulletins* of the London and Cambridge Economic Service, the monthly *Bulletins* of the Institute of Statistics, Oxford, the weekly issues of the *Economist*, and (since January, 1947) its supplement *Records and Statistics*, and the weekly issues of the *Statist*. The *Journal of the Royal Statistical Society* contains current statistical notes as well as full-dress papers, many on matters of current general interest.

Data for the various countries of the world, including the U.K., the Dominions and other Empire countries, are assembled on as comparable a basis as possible by the Statistical Office of the United Nations, previously by the League of Nations. The main statistical publications are the *Monthly Bulletin of*

Statistics and the annual *Statistical Year Book*. The former was published first by the League of Nations and continued without break by the United Nations (since January, 1947, from New York); the latter was issued by the League until the issue covering 1942–4 and is to be continued by the United Nations.

2.2 *National Income and Expenditure*. Official estimates of the total and composition of national income and expenditure are available for years since 1938 and published in the annual White Papers on this topic. The first issue covering the years 1938 and 1940 appeared in 1941 (Cmd. 6261) and the scope was widened considerably in subsequent issues (cf. the issue of 1947, Cmd. 7099). The White Paper of 1945 (Cmd. 6623) contains a detailed account of the basic estimates for the year 1938. The broad outline of the material published can be seen from Table 1 of Appendix I, but a considerable amount of detail is shown, particularly on the analysis of personal expenditure, at current and at 1938 prices, at market prices and at factor cost, and on the expenditures of public authorities and the channels of its financing. An important table (cf. Cmd. 6784, Table 1, National Income and Product) shows the relations between various alternative aggregates which can be obtained: personal income, private income, net national income at factor cost, gross national product at factor cost and at market value. Some information is also provided on the distribution of private incomes (before and after tax) among individuals, on the lines indicated by Table 12 of Appendix I.

The pioneer work on estimates of British national income was done by private investigators, particularly between 1914 and 1939 by Bowley, Stamp and Clark. Three publications on the national income before 1914 and in 1924 are collected together in A. L. Bowley and Josiah Stamp's *Three Studies on the National Income* (1938), and further estimates are included in A. L. Bowley's *Wages and Income since* 1860 (1937). This work owes much to Josiah Stamp's *British Incomes and Property* (1916). Some results of investigations in progress in 1939 are given in a volume edited by A. L. Bowley, *Studies in the National Income*, 1924–38 (1942). Colin Clark's work is seen best in his book, *National Income and Outlay*

(1937), which includes quarterly as well as annual estimates of national income. The main recent work on the national capital is H. Campion's *Public and Private Property in Great Britain* (1939). Amongst more recent investigations, an outstanding study is T. Barna's *Redistribution of Incomes*, 1937 (1945), and the same authority is responsible for quarterly estimates and forecasts of national income published in the *Financial Times* (4th February, 1947).

Studies on international comparisons of national income are made by private investigators, including Colin Clark, *The Conditions of Economic Progress* (1940), J. R. N. Stone (e.g., *Econ. Jour.*, 1942–3), and various authors in the series *Studies in Income and Wealth* published by the National Bureau of Economic Research, New York. The work has been taken up by the League of Nations and the United Nations. A report, including a technical appendix by Stone, has been published as No. 7 in the U.N. series of *Studies and Reports on Statistical Methods* (1947).

2.3 *Finance and Banking.* Much of the information regularly available in this field is of an accounting rather than a statistical nature. On the operations of the central government, the daily Press and such weeklies as the *Economist* publish the weekly *Exchequer Return* on revenue and expenditure, on the floating debt and on other government borrowing. With the Budget each year, the Treasury issues a *Financial Statement* and a set of *Finance Accounts* as House of Commons Papers, giving estimates of revenue and expenditure in comparison with actual receipts and outgoings. Further details of revenue are given in the annual *Reports of the Commissioners of Inland Revenue* and of *H.M. Customs and Excise.* On local finance, the Ministry of Health issues an annual consolidated statement, *Local Government Financial Statistics*, for all local authorities. The annual White Paper on *National Income and Expenditure* contains much supplementary analysis of government operations, particularly on the channels of central government financing.

The Press reproduces the weekly *Bank of England Return*, including information on the note issues, deposits and securities held by the Bank. The Press also publishes certain figures on

the operation of the joint stock banks. There are weekly returns on bank clearings by the *London and Provincial Clearing Houses*, and monthly returns from the *London Clearing Banks*, eleven in number from 1936, showing deposits and other liabilities set against assets in various categories. Some of the data are summarized in Table 6 of Appendix I. An analysis of advances, by groups of industries and other customers, is given in a special quarterly return (since February, 1946) and reproduced in the *Economist* (cf. *Records and Statistics* Supplement, 19th April, 1947).

Information of the foregoing types is summarized over long periods in the *Monthly Digest of Statistics* (following the Bank of England *Statistical Summary*, discontinued in 1939) and in the *Bulletins* of the London and Cambridge Economic Service. The question of international comparability of banking statistics has been considered by the League of Nations, e.g., in the annual *Money and Banking*, and in a technical report issued by the United Nations in *Studies and Reports on Statistical Methods*, No. 8 (1947).

The annual *Report of the Commissioners of Inland Revenue* is the basic source of statistics of incomes brought under review by the income tax authorities (cf. Ninetieth Report, for 1946/7, Cmd. 7362). In particular, income reported under Schedule D relates to profits from businesses, professions, etc. These reports provide the basis for estimates of profits and other elements in the national income and for the estimated distribution of incomes. The data are so technical as to need expert interpretation and, moreover, since 1920 the amount of detail (e.g., in the analysis of Schedule D by industrial groups) of special use to the economist and social scientist has decreased rather than increased.

The most complete data, on a consolidated basis, of profits of individual companies arranged under industrial groups are those published in the *Records and Statistics* supplement to the *Economist*. The data are analysed quarterly to determine the course of profits over time (cf. *Commercial History and Review of* 1947, supplement to the *Economist* of 7th Feb., 1948). The analysis is limited to public companies reporting in the current quarter *and* in the previous year, which makes the data not altogether representative of the profits of all concerns.

2.4 *Population and Vital Statistics.* The main characteristics of the population are determined in the decennial Census of Population, conducted separately for England and Wales and for Scotland, and analysed in great detail in the census volumes. The published data include the geographical and age distribution of the population, occupational and industrial classifications of the working population (gainfully occupied), information on housing, particularly the distribution of households by size and number of rooms occupied, on nationalities, birthplaces and other details. Many of the characteristics are analysed separately for the several administrative areas in a series of county volumes. The census due in 1941 was not taken because of the war, though a partial substitute is provided by the somewhat different, and more limited, statistics of the National Registration at the outbreak of war, as published in *National Register, Statistics of Population on September* 29, 1939 (1944).

Vital statistics are published in great detail and the records, like those of population, go back fairly continuously to the beginning of the nineteenth century. Summary statistics of the numbers of births, marriages and deaths, and the corresponding rates, appear in quarterly returns of the Registrars-General for England and Wales and for Scotland. Full details of deaths and death rates by areas and by a great variety of classifications, and rather less detail of births and marriages are issued in annual reports, the *Registrar-General's Statistical Review of England and Wales*, and the *Annual Report of the Registrar-General for Scotland.* Appendix I, Table 11 shows some data from the *Statistical Review.* Further and more specialized analyses of vital statistics and population trends are made in the *Decennial Supplements*, issued by the Registrars-General following each census of population. These *Supplements* are particularly valuable as sources of data on such topics as life tables and occupational mortality.

The Population Statistics Act of 1938 requires the recording of additional information at the registration of births and deaths. Further tabulations, particularly on fertility, are consequently available since 1938. The new material is described in R. R. Kuczynski's *The New Population Statistics* (National Institute of Economic and Social Research, Occasional Papers

I) and the annual *Statistical Review* now includes a section on fertility and calculations of reproduction rates. This material will be supplemented by the results of the family census conducted in 1946 by the Royal Commission on Population.

Official estimates of the population in administrative areas are made quarterly on the basis of recorded births and deaths and of information on migration. These are published in the quarterly returns of the Registrars-General. Estimates at each mid-year are prepared in more detail, e.g., showing age distributions of the population of the main regions, and issued in the annual publications of the Registrars-General. Projections of the total population for many years ahead, on particular assumptions as regards mortality and fertility, have been made by the Registrars-General in *Current Trend of Population in Great Britain* (Cmd. 6348, 1942). The main work in this field, however, is still left to private investigators. Detailed projections of population are made in G. G. Leybourne's "An Estimate of the Future Population of Great Britain," *Sociological Review* (1934) and in Enid Charles's "The Effect of Present Trends . . .", London and Cambridge Economic Service *Special Memorandum* No. 40 (1935). Reference can also be made to A. M. Carr-Saunders's *World Population* (1936), and to D. V. Glass's *Population Policies and Movements in Europe* (1940).

Demographic statistics, including analyses of past population growth and future projections, are extensively developed on an international basis by the International Statistical Institute and the League of Nations. A recent publication by the League is *The Future Population of Europe and the Soviet Union, Population Projections*, 1940–1970 (1944). The work is being continued by the Population and Statistical Offices of the United Nations and the publication of a Demographic Year Book is planned.

2.5 *Manpower and Labour Statistics.* Information on unemployment, primarily among workers insured against unemployment, is published in detail by industries and by regions in the monthly Ministry of Labour *Gazette*, with longer runs of figures up to 1936 in the *Abstract of Labour Statistics*. Annual estimates (each July) of the total numbers

of insured workers, and hence of employment in insured trades, are published in the November issues of the *Gazette*. The *Gazette* also includes tabulations of the duration of unemployment and of other details of the incidence of unemployment. There are early data based on returns from trade unions, published in the *Gazette* up to 1927, but more recent figures depend on records of unemployment insurance, varying in scope from time to time since the major Acts of 1920–1.

Direct estimates of the total working population and of total employment in groups of industries are given monthly in the *Gazette*, and are summarized, with some additional analyses in the *Monthly Digest of Statistics*. These data are taken back to 1939 and their general nature is indicated in Table 2 of Appendix I. They include estimates of the numbers in the armed Forces, and of those directly employed on export orders. This information provides the basis for estimates of the future distributions of total manpower, for example, as given officially in the Ministry of Labour *Gazette*, May, 1947 and in *Economic Survey for* 1948 (Cmd. 7344), and by such private investigators as T. Barna in "A Manpower Budget for 1950," *Bulletin* of the London and Cambridge Economic Service (October, 1945).

Information on standard wage rates and hours of work is published regularly in the Ministry of Labour *Gazette* and collected in the Ministry's handbook, *Time Rates of Wages and Hours of Labour* (cf. issue of September, 1947). Average weekly earnings and average hours worked are obtained for manufacturing and for some other industries through special inquiries undertaken by the Ministry of Labour at intervals since 1906 and published in the *Gazette*. Since 1946, inquiries have been made twice yearly in April and October. Other data on earnings are available in the census of production, now to be taken annually.

The topics of wages, earnings and hours are studied by many private investigators, and indeed much of the pioneer work was done by them. Historical estimates are summarized in A. L. Bowley's *Wages and Income since* 1860 (1937). The same authority is responsible for current surveys and records of wage rates in a representative list of occupations appearing periodically in the *Bulletins* and *Special Memoranda* of the

London and Cambridge Economic Service. Other studies appear in such publications as the *Bulletin* of the Institute of Statistics, Oxford (for example, an article by J. L. Nicholson in 1946).

Details of trade disputes in progress and settled appear regularly in the Ministry of Labour *Gazette* and are summarized in the annual *Statistical Abstract*. The *Abstract* is also the most convenient first reference for statistics of industrial accidents and workmen's compensation.

2.6 *Production and Wholesale Prices.* An annual census of agriculture, relating to all holdings of more than one acre, is taken to give the acreage under various crops and grass and the numbers of livestock. A census of livestock is also available quarterly since 1940. The crop reporters of the Ministry of Agriculture make yearly estimates of crop yields and hence of quantities harvested. The main source of these statistics is the annual *Agricultural Statistics*, discontinued in 1939 and resumed in 1947, and some of the data are included in Appendix I, Table 8. The Ministry of Food provides monthly statistics of the movement of the main crops off farms and of animals purchased for slaughter, statistics which are summarized in the *Monthly Digest of Statistics.*

The Ministry of Fuel and Power issues monthly (or quarterly) and annual statistics of coal production, consumption and stocks, of employment and output per manshift, of costs of production and of other details of coal-mining. The sources are Press releases, quarterly returns issued as Command Papers and the annual Ministry of Fuel and Power *Statistical Digest*. The last was first issued in 1944 (*Statistical Digest from* 1938, Cmd. 6538) and takes the place of the *Annual Reports* of the Mines Department before 1939. Appendix I, Table 13 is taken from the *Digest* for the year 1945. The same sources give data on coke, gas and electricity production and on supplies of petroleum.

Regular information on production, consumption and stocks of raw materials is available in the publications of the industries concerned, e.g., the monthly *Statistical Bulletin* of the British Iron and Steel Federation and the quarterly *Statistical Supplement* to the Cotton Board's *Trade Letter*.

The most convenient general source is the *Monthly Digest of Statistics* which includes quantity data provided by the Ministry of Supply, the Board of Trade and the Ministry of Works. The same source gives production data (usually by quantity but sometimes only by value) for many manufactured goods. These are from the Ministry of Supply for capital goods such as shipbuilding, motor vehicles, machine tools, electric motors, stationary engines and various types of machinery, and from the Board of Trade and Ministry of Food for such consumers' goods as footwear, hosiery, pottery, electrical appliances, foodstuffs, drink and tobacco. The Board of Trade publishes regular statistics of supplies of consumers' goods in the home civilian market, i.e., domestic production for the home civilian market together with imports (if any). These appear in the Board of Trade *Journal* and are summarized in the *Monthly Digest of Statistics*. The *Housing Returns* for England and Wales, and for Scotland, issued as Command Papers monthly, are the source of data on construction and completions of houses by local authorities and private builders.

A census of production was taken at intervals up to 1935 and was then intended to occur every five years. Under the Statistics of Trade Act of 1947, the census is to be conducted annually; a partial census was taken in respect of the year 1946 and a full census for 1948 is in preparation. The results of each census, in a series of census volumes issued by the Board of Trade, comprise statistics of annual output (by quantity) and of the value of gross and net output of manufacturing, building, mining and some other trades, with related data on employment, earnings, costs and capital equipment. The scope of the material collected is being extended in post-war census inquiries. A census of the distributive trades and services is in preparation, to be taken every five years under the Statistics of Trade Act.

A great variety of wholesale price quotations is available regularly for agricultural produce, foodstuffs and materials, but only scattered quotations are to be found for manufactured goods. Agricultural prices of cereals, farm crops and livestock are collected together by the Ministry of Agriculture in the annual *Agricultural Statistics*, and one series from this source is given in Table 7 of Appendix I. A selection of more than

100 wholesale prices of agricultural produce, foodstuffs and materials appears monthly in the *Records and Statistics* supplement to the *Economist* and another selection was published monthly up to 1939 in the Board of Trade *Journal*. The *Journal of the Royal Statistical Society* includes each year a long run of the annual averages of the prices of forty-five selected items as used in the construction of the *Statist* index of wholesale prices. These series go back for more than 100 years. Even longer runs of particular price quotations of foodstuffs and materials are collected in W. H. Beveridge's and others *Prices and Wages in England* (Vol. I, 1939).

2.7 *Trade and Transport.* Tabulations of the recorded merchandise trade of the U.K. with overseas countries are published monthly by the Board of Trade in the *Accounts Relating to Trade and Navigation of the U.K.*, and yearly in the *Annual Statement of the Trade of the U.K.* The monthly tables show a detailed commodity classification of imports, exports and re-exports by value and usually by quantity. There is a subsidiary classification by country of origin or destination varying in detail according to the importance of the commodity group. Total trade with each country is given monthly in the *Accounts*. The annual volumes show trade each year in much greater detail by commodities, countries and ports. Appendix I, Table 3 illustrates the type of data available by commodities.

The balance of international payments on current account includes "invisible" items of receipts and payments, as well as merchandise trade. Annual estimates of the balance of payments up to those for the year 1938 appear in the Board of Trade *Journal*. Post-war estimates are on a different basis, using data from exchange control and other sources. They appear (e.g.) in *U.K. Balance of Payments*, 1946 *to* 1948 (Cmd. 7520) and in *Economic Survey for* 1948 (Cmd. 7344).

The League of Nations has issued numerous studies and collections of international statistics of external trade and balances of payments on a comparable basis, and the work is being continued by the United Nations. There are annual volumes up to 1938 in the League's series on *International Trade Statistics*, *International Trade in Certain Raw Materials*

and Foodstuffs and *Balances of Payments*. Recent publications include *Europe's Trade* (1941) and *The Network of World Trade* (1942) and a technical report on the content of a balance of payments issued by the United Nations in *Studies and Reports on Statistical Methods*, No. 9 (1947).

Statistics on transport and communications are voluminous for railways, adequate for shipping and the post office, in process of reformation for civil aviation and very scanty for road transport. The sources of data on the railways are the monthly *Railway Statistics* and the annual *Railway Returns* of the Ministry of Transport. The changing nature and efficiency of railway operation is seen by relating such a figure as ton-miles of freight carried to other figures such as tonnage of freight (to give average length of haul) or number of engine hours operated. Only summary tables are available for the war years 1939–45. The main shipping statistics relate to tonnage of vessels entering and clearing U.K. Ports in foreign and coastal trade, as published monthly in the *Accounts Relating to Trade and Navigation of the U.K.*, and yearly in the *Annual Statement of the Navigation and Shipping of the U.K.* Information on Post Office operations in the mail, telephone and telegraph services is given in the *Commercial Accounts* issued annually as a House of Commons Paper. Details of traffic accidents are readily available in publications such as the Annual Report to the Minister of Transport on railway accidents and the Ministry of Transport's Reports on road accidents.

2.8 *Consumption and Retail Prices.* Annual statistics of the value of personal expenditure on consumption, for the whole range of consumers' goods and services, form part of the material in the White Papers on *National Income and Expenditure*, the valuation being both at market prices and at factor cost (i.e., adjusted for indirect taxes and subsidies). Quarterly estimates are now appearing in the *Monthly Digest of Statistics*. The broad method of estimation is a valuation of the quantities of goods flowing through distributive channels to consumers and of the services currently rendered.

A different set of data on purchases of consumers' goods is provided by the monthly statistics in the Board of Trade

Journal on the value of retail sales of the main categories of merchandise for a group of retail outlets (mainly the larger department stores, multiple shops and co-operatives). The figures are expressed as sales per week (per selling day before 1947) in percentage of the sales of the same group of outlets in the corresponding month of the previous year. Such percentages can be "chained" together to give a continuous run of figures. Similar data are available in the *Journal* since 1946 for some groups of smaller (independent) retailers on the basis of monthly sample inquiries as described in the issue of the *Journal* of 19th October, 1946.

Statistics of the distribution of family expenditure are available as the result of family budget inquiries, as for example that of Table 10 of Appendix I. Official budget inquiries were taken very infrequently before 1939, and private inquiries more frequently, but for limited groups of families. For early collections of family budgets, see R. G. D. Allen's and A. L. Bowley's *Family Expenditure* (1935). A large budget inquiry was undertaken by the Ministry of Labour in 1937–8 for working-class families (urban and rural) and the main results appear in the Ministry of Labour *Gazette*, December, 1940—February, 1941. An unofficial collection of middle-class family budgets was made in 1938–9, see Philip Massey's "The Expenditure of 1,360 British Middle-class Households in 1938–9," *Jour. Roy. Stat. Soc.* (1942). Many budget inquiries were made during the War of 1939–45, but no results are published. The Ministry of Labour is now considering the means of collecting family budgets at regular intervals (see *Interim Report of the Cost of Living Advisory Committee*, Cmd. 7077, 1947).

Actual retail prices are difficult to measure, partly because of variations in the quality of the commodity priced. They are often quite impossible to obtain in practice. Apart from the retail prices of bread and a few other items, information before 1914 is very scanty and what exists is summarized in A. L. Bowley's *Wages and Income since* 1860 (1937). Subsequently and until June, 1947, the Ministry of Labour *Gazette* shows monthly the actual (average) prices of basic food items purchased at retail by working-class families (see Appendix I, Table 4). Percentage changes from month to

month are also shown for certain groups of goods and services —food, rent and rates, clothing, fuel and light, and miscellaneous items. New methods of collecting retail prices were introduced in June, 1947 (see Ministry of Labour *Gazette*, August, 1947).

2.9 *Social Statistics.* Many general surveys of social conditions are available in published form for particular areas and at particular dates. All such surveys were privately organized and generally cover a wide range of topics—industrial distribution and development; income, expenditure and poverty; housing, health, education and other public services; entertainment and recreation; and so on. The pioneer work was done by Charles Booth in London (*Life and Labour of the People in London*, nine volumes, 1892–7) and by Seebohm Rowntree in York (*Poverty, A Study of Town Life*, 1901). Later surveys on an extensive scale include *The New Survey of London Life and Labour* (nine volumes, 1930–5), *The Social Survey of Merseyside* (three volumes, 1934) and Seebohm Rowntree's *Poverty and Progress* (1941). Appendix I, Table 14 is taken from the first of these three surveys. More recently, surveys have been made to aid in the reconstruction and development of particular areas. J. H. Forshaw's and L. P. Abercrombie's *County of London Plan* (1943), and Janet Glaisyer's and others *County Town* (1946) are outstanding examples of surveys sponsored by civic authorities.

Amongst official publications, those in the fields of education, health, pensions and relief are often of administrative rather than of general interest. They include the annual reports of the Ministries of Education, Health and Pensions, the report of the Chief Medical Officer of the Ministry of Health, the report of the Commissioners of H.M. Customs and Excise (on non-contributory old-age pensions) and the quarterly returns of the Ministry of Health on *Persons in Receipt of Poor Relief*. There are, however, some documents of wider scope or of more particular interest. The University Grants Committee issues annual returns including data on the development of university education. The Medical Research Council has issued occasional reports from its Industrial Health Research Board. In the field of National Insurance, there are

A. Watson's "National Health Insurance, A Statistical Review," *Jour. Roy. Stat. Soc.* (1927) and some official reports issued in connection with the National Insurance Bill, 1946. Finally, the reports of the Assistance Board contain much information on the position of the aged members of the community.

Statistics of crime and related matters generally need expert interpretation. The main sources of data are the Home Office *Criminal Statistics* and the Lord Chancellor's Department *Civil Judicial Statistics*, for England and Wales, and corresponding annual returns for Scotland. In specialized fields there are the *General Annual Reports* on Bankruptcy and on Companies by the Board of Trade, for data on insolvency, the yearly House of Commons Papers on *Offences Relating to Motor Vehicles*, and the annual Command Papers on *Licensing Statistics*, including convictions for drunkenness.

A wide range of statistics is included in the *Reports of the Chief Registrar of Friendly Societies* and in the statistical summaries issued in advance of these reports. These are the sources of data on the membership and other details of co-operative societies, trade unions and building societies, amongst other types of friendly societies.

Practically all the official documents mentioned in this chapter ceased to be issued in the period 1939–45. Publication has generally been resumed (by 1948) though often in a different form. Extensive changes are being made in the nature and scope of official statistics, particularly as a consequence of the nationalization of certain industries and services, and of the passing of the National Insurance and Health Acts. For example, the National Coal Board now (1948) issues a quarterly *Statistical Statement*, and the British Transport Commission puts out every four weeks a publication on *Transport Statistics*. On the other hand, it is not known as yet (in 1948) how estimates will be presented of the numbers of persons insured under the new scheme of National Insurance (cf. 2.5. above).

CHAPTER III

GRAPHS AND DIAGRAMS

3.1 *Objects of Graphical Representation.* Statistical data are first condensed and presented in tabular form. Even so, they cannot be left to tell their own story; their message must be brought out more clearly. Methods must be devised to isolate and describe the general trends and the significant variations in the data. Other methods are needed to show up the relations underlying the material, to correlate one figure with another. All this implies a further summarization of the data, a drastic condensation into a small number of compact figures and relations. This is the job of statistical methods.

One very simple but effective form of statistical analysis is to represent the tabular data by drawing graphs or diagrams. If made with skill and care in avoiding bias, a diagram will show the data in a graphical form in which the salient features leap to the eye. The risk is that diagrams can be misleading when drawn by the unskilled and they can be very dangerous tools in unscrupulous hands. In this chapter we shall see how a good graph or diagram should be drawn and we shall discover some of the errors to avoid. As with tabulation, however, skill in constructing diagrams is only acquired after long experience. The main point can be easily made; a graph or diagram should be clear and simple since it adds nothing to our understanding if it does not show up the trends and relations of our data more obviously than in the original tables. A chart is meant to "help out" in drawing broad conclusions from a table which may be quite complicated. Inevitably the graph or diagram is less exact and shows less detail than the table; it is a step in the constant process of summarizing data. This must not be overdone. It is only too easy to simplify so drastically as to be misleading.

3.2 *The Graph of a Time Series.* As Tables 5, 6 and 7 (Appendix I) show, the simplest of all statistical tables exhibit

time series. We have seen, however, that the form of such a table may be simple but it is often necessary to add complicated explanatory notes. Circumstances change as time goes on and there is always the need to ensure comparability from one period to another. Further, as we shall show later, the handling of time series raises problems of great complexity. For the moment, we shall ignore such complications and concentrate on the task of representing a given series in graphical form.

A time series is a single variable quantity given at successive points or intervals of time, as the series of employment percentages shown yearly (each July) in Table 6. In this instance, comparability over time is by no means perfect and there is one definite break in the series in 1938. A time graph of such a series is basically a very simple concept. A piece of squared paper is taken, a horizontal and a vertical axis are drawn, time is measured along one axis (usually the horizontal) and the variable along the other. Most of the difficulty lies in the choice of appropriate scales for time, and particularly for the variable.

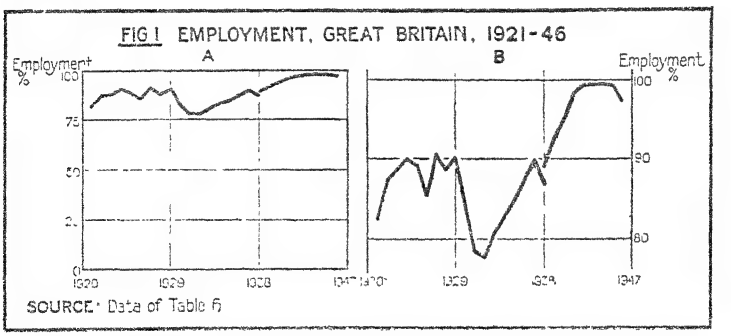


Fig. 1A shows a time graph of the employment series, in which the vertical scale ranges from zero at the base line of the graph to 100 at the top. The points on the graph are joined by straight lines for convenience and their heights above the base line show the variation in the employment percentage. An increase from (say) 80 to 90 in the percentage is represented by a rise of one-eighth in the height of points of the graph. The disadvantage of this representation is that the up and down movement is not large, being confined to the

top quarter of the graph, leaving the other three-quarters blank. If we are to see the variations more clearly—or if we are pressed by the publisher to eliminate the blank space—then we must enlarge the scale of the variable and start it from a point above zero. This is done in Fig. 1B. In this diagram we must avoid comparing heights of points above any line at the bottom of the graph, for their heights now have only an arbitrary meaning. If the bottom line is drawn at the 75 per cent level, then an increase from 80 to 90 in the employment percentage is represented by points rising from five to fifteen units above the base line, i.e., the second point is three times the height of the first instead of one-eighth higher. It follows that, in a graph such as Fig. 1B, great care must be taken to indicate that there is no zero base line in the graph. This can be done by breaking the vertical axis towards the bottom and by indicating the break clearly. Better still, we can draw no line at all at the bottom of the graph, inserting instead one or more horizontal lines to serve as guides, say at the levels of 80 per cent, 90 per cent, and 100 per cent.

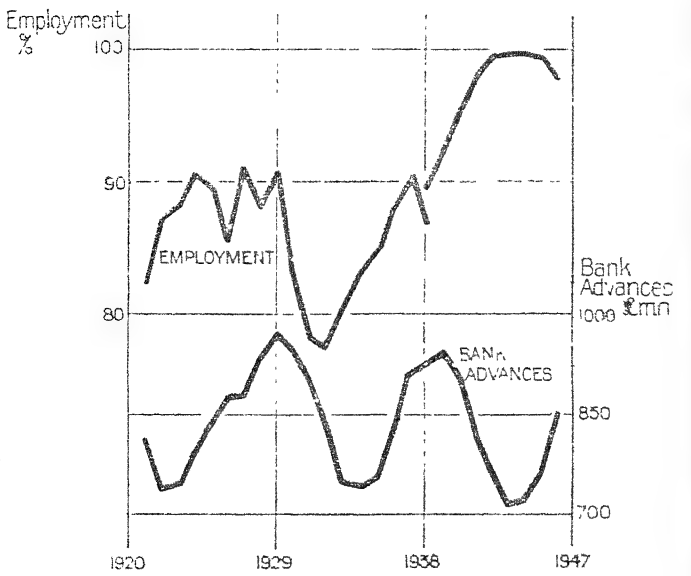
Whatever is done about the base line, there is always a decision to be taken on the relative scales of the horizontal and vertical axes. The rectangle which encloses the graph can have its horizontal side longer or its vertical side. The choice depends partly on the general appearance of the graph. Mainly, however, it is a matter of the amplitude required to be shown in the variable. The vertical side of the rectangle must not be so short as to make variations up and down insignificant, nor so long as to make the oscillations violent. A graph of the monthly series of egg prices in Table 7 would illustrate this point very well. There is here no difficulty about the zero base line for a vertical scale from zero to thirty shillings is adequate. The risk is that, by making the vertical side long relative to the horizontal, the graph will oscillate so much that we get no real picture of month to month variations.

3.3 *Graphical Comparison of Time Series.* A comparison of two or more time series is easily made when the variables of the series are of the same nature and given in the same units. When a suitable graph is drawn for one series, the others can be added to the graph with exactly the same scales. For example,

the employment percentage for Great Britain shown in Table 6 may be given separately for London and other regions. Such regional series can be added to Fig. 1 and comparisons can then be made between the variations of employment in various regions of the country. If there is a zero base line in the graph, as in Fig. 1A, actual values of the different series are compared. If the base line is omitted, as in Fig. 1B, a comparison of actual values is not possible but variations (up and down) in one series can still be compared with those in another. The vertical scale is the same for all series and rises or falls in one series compare directly with rises or falls in another.

Complications arise in the graphical comparison of time series when the units in which the variables are measured are not the same. For example, the course of bank advances in

FIG.2 EMPLOYMENT AND BANK ADVANCES, GREAT BRITAIN, 1921-46



SOURCE: Data of Table 6

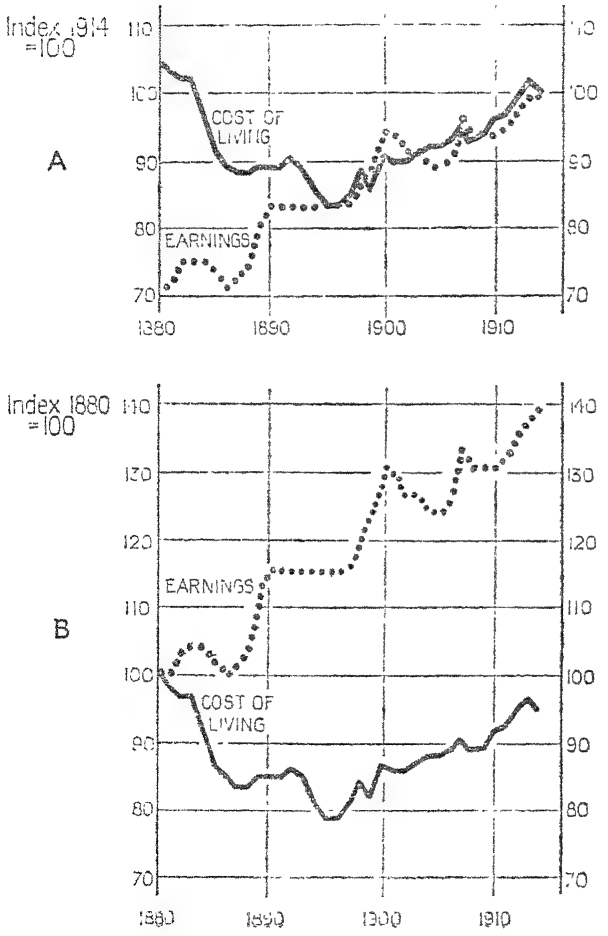
£ millions is to be compared with that of employment as a percentage (data of Table 6). The two series are in different units and a separate choice of scales must be made for each. There is no natural relation between the units; £10 millions of bank advances cannot be said to correspond to 1 per cent of unemployment or to any other such figure. In such a case, a graph of the two series does not permit a comparison between variations, e.g., between the amplitude of the fluctuations of one series and that of the other. The actual variations shown depend entirely on the scales chosen. The graphical comparison is therefore limited but it can still be useful. It will show when the series go up and down together and when they go in opposite directions. In particular, it will indicate whether there is any relation between the turning points, the peaks and the troughs, of the series.

Fig. 2 illustrates the method of graphing two series with different scales. Here the scales are so chosen that the amplitudes of variation of the series are roughly equal, and so placed on the vertical axis that one series lies completely above the other. The graph would not be so good if the scales were so chosen that one series fluctuated wildly and the other little, or so placed that the series criss-crossed each other in a confusing way.

There is one device, used in plotting several time series on one graph, which may appear to overcome the difficulty about choice of scales. One date is selected as a base point for both series. The value of the variable at the base date is written as 100 in each series and the values at any other date are expressed in percentage of the base value. The series of earnings and cost of living in Table 5 are adjusted in this way with the value in 1914 as 100. The original units for measuring earnings and cost of living are discarded and the series are each in percentage form and so directly comparable. A graph showing both series can be drawn immediately, as in Fig. 3A.

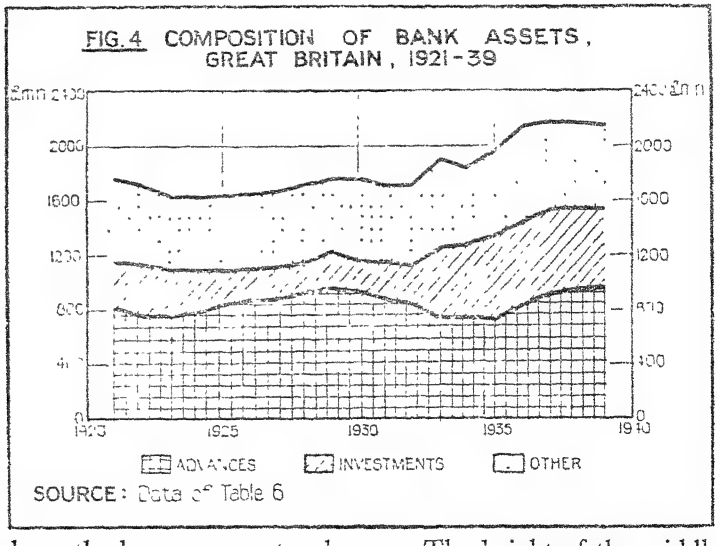
This is, however, no solution of the difficulty. The device merely shifts the arbitrary decision from the choice of scales for different units, to the choice of a date to be taken as 100. The selection of a different date as base changes the whole series of values (percentages) and the resulting graph has a

FIG.3 EARNINGS AND COST OF LIVING, U.K., 1880-1914



different appearance. A comparison of Fig. 3A and Fig. 3B shows this. The series used are the same except that 1914 is the base year in Fig. 3A and 1880 in Fig. 3B. The up and down variation in earnings shows up much more, relative to that in the cost of living, in Fig. 3B, a consequence of the fact that the choice of 1880 as base date implies a choice of a larger scale for earnings.

A different method of graphing is possible when the series to be plotted are constituents of the same total. In Table 6, bank advances and investments are two assets of the banking system which with other assets make up total assets, equal to deposits. Having the same scales, all these series could be plotted on the same graph, each to be related to the same zero base line. However, since they add to a total in each year, they can be graphed cumulatively, one piled on top of the other, as in Fig. 4. The height of the lowest set of plotted points



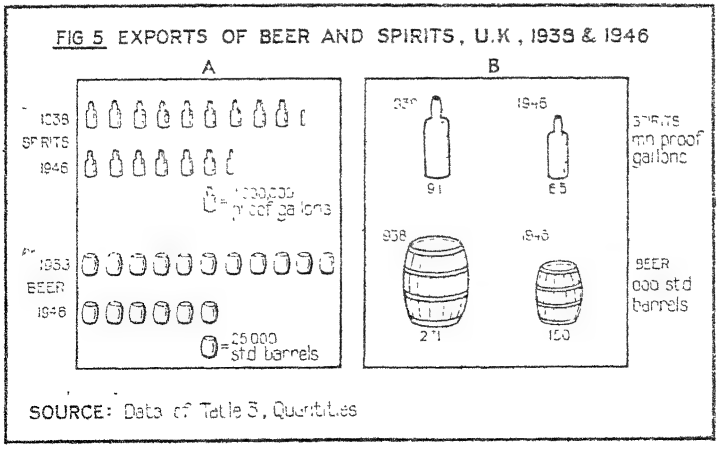
above the base represents advances. The height of the middle set is advances plus investments, so that the difference between the middle and lowest tier of points represents investments. Similarly, other assets are given by the difference between the highest and the middle tiers, the former showing total

assets (equals deposits) when related to the base line.

Attention should be paid to the "finishing off" of graphs like Figs. 1–4 to give them a neat and attractive appearance. There are various styles in which graphs can be produced and little to choose between many of them. The diagrams here are presented with an outer border, surrounding the rectangle of the graph itself and in this border is placed the description of the scales, the heading and the footnotes appropriate to the graph. Whatever style is adopted, the graph, like the table from which it derives, should show a description of contents in the heading and a specification of the sources of the data in a footnote.

3.4 *Pictorial Diagrams.* It is increasingly fashionable to represent many types of tables, and particularly those in which the classification is by attributes, by means of pictorial diagrams. These can be extremely attractive and, if well and correctly drawn, they attract attention without being misleading.

Some of the data on exports given in Table 3 (Appendix I) is presented pictorially in Figs. 5 and 6. Alternative uses of representative symbols are shown in Fig. 5, in this instance bottles for spirits and barrels for beer exported in two years. In Fig. 5A, rows of small symbols show quantities exported,



each bottle representing one million proof gallons of spirits and each barrel 25,000 standard barrels of beer. In Fig. 5B, the same quantities are shown by single symbols which vary in size according to the quantity. Size must be determined here by the area of the symbol since this is what strikes the eye. Since length of row is easier to judge than an irregular area, the first method is to be preferred to the second.

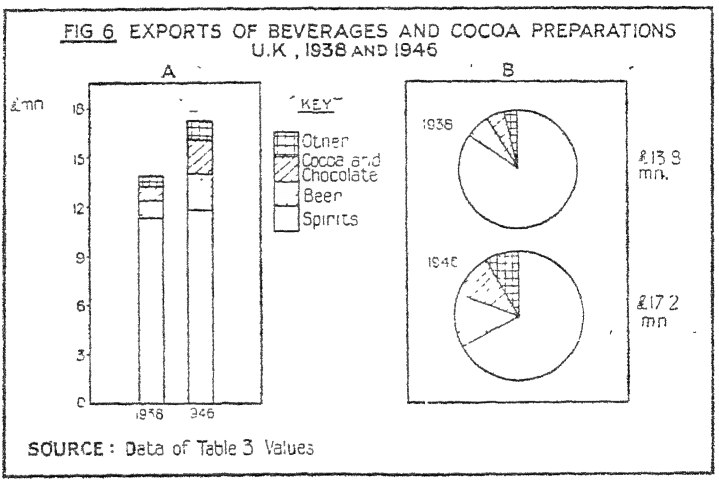
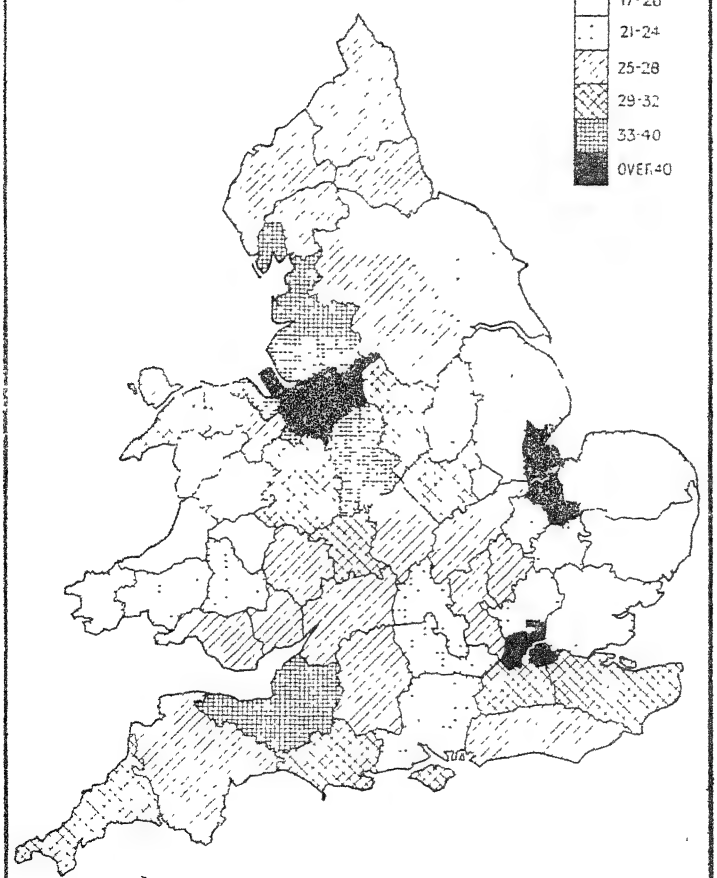


Fig. 6 is designed to show how a total is composed of various constituents, in this case, the total value of exports of beverages and cocoa preparations in each of two years. Two forms are given, again depending on the judgment of length and area, respectively. In Fig. 6A, thin rectangles are drawn with lengths proportional to total values represented, and each is divided into parts which show by their lengths the values of exports of spirits, beer, cocoa preparations and other items. The alternative representation in Fig. 6B uses what is adequately described as a "pie" diagram. For each year, a circle or "pie" is drawn with area representing total value of exports, i.e. with radius proportional to the square root of the value. The "pie" is then cut up into segments with areas representing the values of the various items exported. The segments are most easily determined by taking the angles

FIG 7 AGRICULTURAL RENTS BY COUNTIES, ENGLAND AND WALES, 1941.



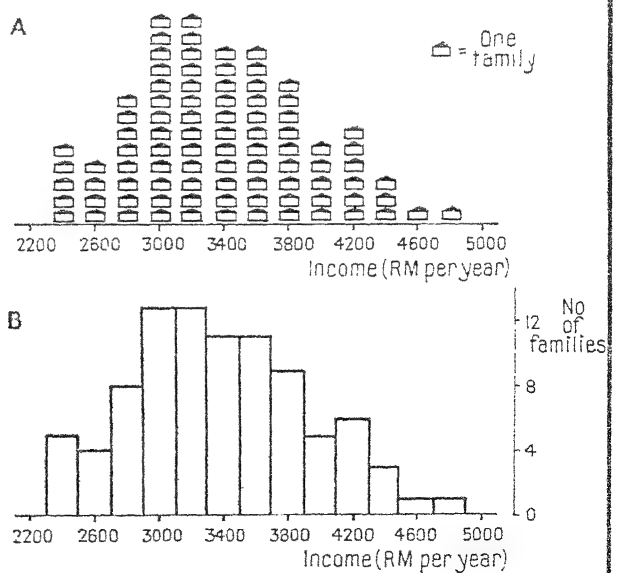
SOURCE: Data from Table 8 (extended to all counties), from National Farm Survey of England and Wales (1946)

at the centre in proportion to the constituent values in percentage of the total. The chart using rectangles is the more easily read but the "pie" diagram is quite accurate and clearly has much to recommend it as a very striking presentation of the data.

Another pictorial diagram is the cartograph which shows the geographical distribution of a given character. The data on agricultural rents in Appendix I, Table 8 (extended to all countries in England and Wales) are shown pictorially in the cartograph of Fig. 7, a good representation of the geographical variation of agricultural rents.

3.5 *Diagrams of Frequency Distributions.* It remains to show

FIG. 8 DISTRIBUTION OF FAMILIES BY INCOME, HAMBURG AND BREMEN, 1927-28



SOURCE: Data of Table 106

how we can construct diagrams to represent a frequency distribution in which items are classified according to ranges of a variable character. As a simple example, in which the classes are all of the same length, we can take the distribution of ninety families by income shown in Appendix I, Table 10B.

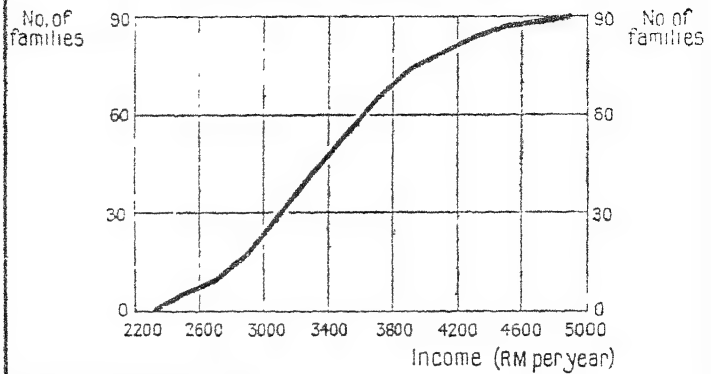
First, a type of pictorial diagram can be drawn as in Fig. 8A. Here a small symbol representing a house is taken for each family and in each income class the appropriate number of symbols is piled up to indicate the number of families in the class. So, five symbols are piled for the range 2,300–2,499 RM, four for the range 2,500–2,699 RM, and so on.

A less picturesque but more useful diagram can then be constructed as in Fig. 8B. A horizontal axis is drawn and marked off in ranges of income according to the given classes. On each segment of the axis, a vertical rectangle is drawn with width equal to that of the segment and height proportional to the number of families in the class. Since the segments are all of equal length, the areas of these rectangles are also proportional to the numbers of families. The result is a *block diagram* or histogram which indicates by rectangular *areas* the distribution of families according to income. An alternative diagram, sometimes used, is the frequency polygon obtained by joining by a straight line the mid-point of the top of one rectangle to that of the adjacent rectangle.

3.6 *Cumulative Diagrams.* Any frequency distribution can be represented in an alternative form in which the entries in the original table are cumulated downwards. The *cumulative table* of the distribution of ninety families by income is shown in Table 10C; it is obtained and interpreted as follows. First, there are five families with income under 2,500 RM, the number that appears in the class 2,300–2,499 RM. Next, there are nine families with income under 2,700 RM, the first five plus four families in the class 2,500–2,699 RM; and so on.

This table can be represented graphically in very much the same way as a time series (Fig. 9). Income from 2,300 to 4,900 RM is marked on a horizontal axis and numbers of families from 0 to 90 on a vertical axis. Points are plotted over the successive points (2,300, 2,500, 2,700 . . .) which separate the given income classes on the horizontal axis, each point

FIG. 9 CUMULATIVE DISTRIBUTION OF FAMILIES BY INCOME, HAMBURG AND BREMEN, 1927-28



SOURCE: Data of Table 10C

having height (0, 5, 9 . . .) obtained from the cumulative table. The points are then joined, for convenience, by straight lines and the result is an ogive or *cumulative diagram.* The interpretation is a very useful one; the height of each point on the diagram represents (according to the vertical scale) the number of families with income below the level shown by the corresponding point on the horizontal scale.

The convention of joining points by straight lines has a very definite significance. It implies, in fact, that the families within any given income class are distributed uniformly as regards income over the class. If we have only the frequency table and not the original data, we do not know the actual incomes of these families; hence we *assume* conventionally that their incomes are spaced uniformly between the limits of the class. The cumulative diagram, with this assumption, provides a very simple method of interpolation, as do most graphs and diagrams. We know, for example, that there are nine families with incomes less than 2,700 RM and seventeen with incomes less than 2,900 RM. We do not know how many families have incomes less than 2,800 RM. But we can estimate this number from the cumulative diagram by reading off the height of the point on the diagram above the point corresponding to 2,800 RM on the horizontal scale. The number so estimated is thirteen families.

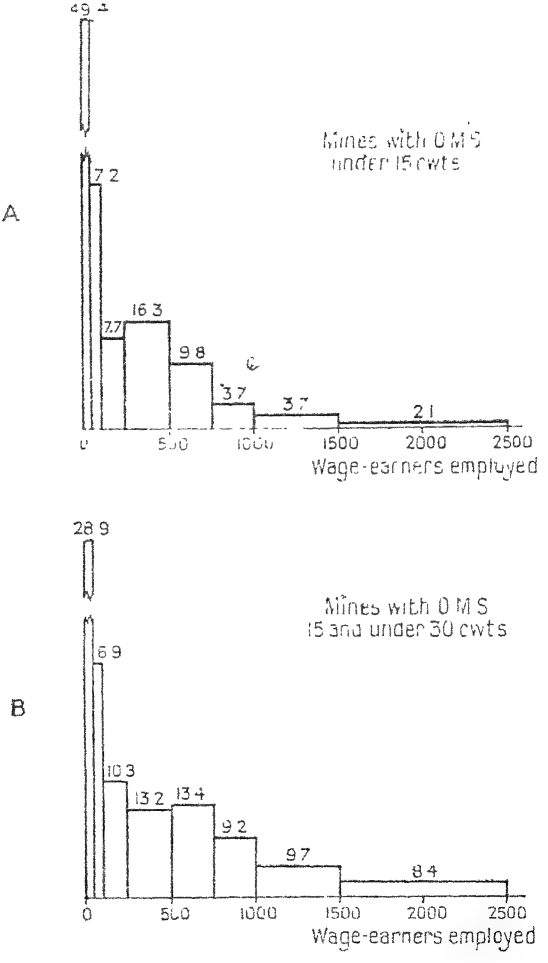
3.7 *Graphical Comparison of Frequency Distributions.* Two frequency distributions can be compared graphically by drawing a block diagram for each, placed side by side or one above the other. For example, to Fig. 8B can be added another block diagram representing the income distribution of a second group of families (e.g., at a different date or place). There is one difficulty arising from the fact that the total number of items may differ from one distribution to the other, affecting the areas of the rectangles in the diagrams. This is overcome by the simple device of reducing each distribution to percentage form in which the number of items in a class is written as a percentage of the total number of items. Any class of one distribution can then be compared directly with the corresponding class of the other, and similarly for the rectangles of the block diagrams.

It has been assumed so far that the classes of a frequency distribution are of equal length, so that the numbers in the classes can be taken equally well for heights and for areas of rectangles in a block diagram. This is not always the case, and indeed it is often impossible in practice. The distribution of coal mines by size (Appendix I, Table 13) shows many small pits and a few of large size. An even classification would give *either* almost all mines in the first few classes (if they are broad) *or* a large number of narrow classes with many nil entries.

Uneven grading of a distribution gives rise to difficulties in the plotting of a block diagram. Two distributions derived from Table 13 are represented as block diagrams in Fig. 10. The distributions relate to mines with differing output per manshift, defined as annual output of coal from the mine divided by the number of manshifts worked in the year. The first step is to reduce each distribution to percentage form as in cols. (2) and (5) below. Next, since the *areas* of the rectangles in Fig. 10 are to be proportional to the percentages of mines in the size groups, the uneven grading means that the *heights* of the rectangles are *not* given by these percentages. The problem is handled by a procedure which turns the entries in cols. (2) and (5) into entries per a fixed length of class (here taken as 250 wage-earners). The results are shown in cols. (3) and (6).

FIG.10 DISTRIBUTION OF COAL-MINES BY SIZE, GREAT BRITAIN, 1945

Figures shown are areas of blocks (% of mines in each group)



SOURCE: Data of Table 13

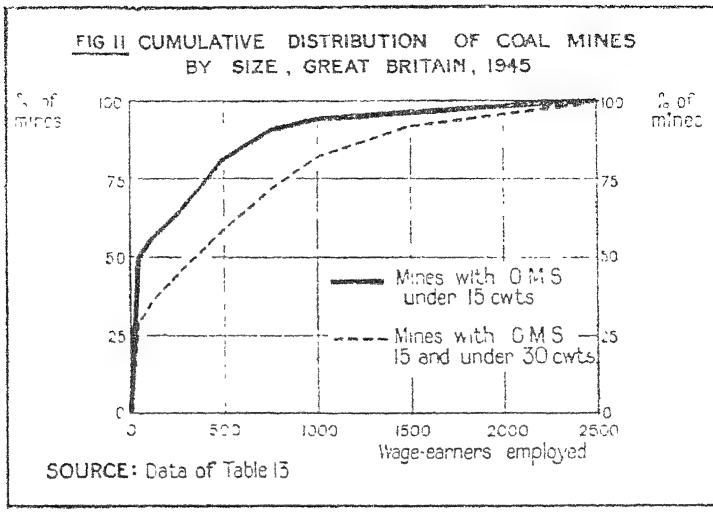
No. of wage-earners employed	Mines with O.M.S. under 15 cwts.			Mines with O.M.S. 15 and under 30 cwts.		
	No.	%	% per standard class[2]	No.	%	% per standard class[2]
	(1)	(2)	(3)	(4)	(5)	(6)
1–49	212	49·4	247·0	301	28·9	144·5
50–99	31	7·2	36·0	72	6·9	34·5
100–249	33	7·7	12·8	107	10·3	17·2
250–499	70	16·3	16·3	137	13·2	13·2
500–749	42	9·8	9·8	140	13·4	13·4
750–999	16	3·7	3·7	96	9·2	9·2
1,000–1,499	16	3·7	1·9	101	9·7	4·9
1,500 and over[1]	9	2·1	0·5	87	8·4	2·1
Total	429	100·0		1,041	100·0	

[1]Assumed to be 1,500–2,499. [2]Of 250 wage-earners each.

The figures in cols. (3) and (6) provide the real comparison of entries in various size groups and they can be used to fix the heights of different rectangles. When multiplied by the "length" of the class they give products proportional to the original percentages, i.e., as heights they give the proper areas of the rectangles. Fig. 10 is then drawn with rectangle heights from cols. (3) and (6), and the figures placed at the top of the rectangles are the original percentages of cols. (2) and (5). Notice that the first rectangle in each diagram is so tall that it cannot be contained within the figure (without reducing the other rectangles to insignificance). A convenient device is adopted whereby this rectangle is broken to indicate that it runs off the paper. Notice also that the last class of the frequency distributions is "open" (1,500 wage-earners and over) and that, for the purpose of this graphical representation, it is *assumed* to be closed (1,500–2,499 wage-earners).

The cumulative diagrams of the same data can be drawn without difficulty as in Fig. 11. The uneven classes are now represented simply as points unequally spaced on the horizontal axis with the points on the graph spaced unevenly to correspond. The percentage distributions, i.e., cols. (2) and (5) cumulated, are again used so that the two cumulative diagrams appear on the same graph and end at the same point (100 per cent). This is necessary for a valid comparison.

The purpose of any statistical presentation is to make some comparison, e.g., between one frequency distribution and another. Figs. 10 and 11 permit a comparison of the size of mines with small output per manshift (under 15 cwts.) with



that of mines with larger output per manshift (15 and under 30 cwts.). The fact that the mines with small O.M.S. are generally smaller in size than those with larger O.M.S. is clearly seen in the cumulative diagram. The cumulative line for the mines with small O.M.S. lies always above that for the other, i.e., there are more mines below any named size in the group with small O.M.S.

3.8 *Ratio Scales.* The graphs and diagrams drawn so far use "natural" scales which space out evenly the divisions corresponding to numbers increasing uniformly in size. A different type of scale can be devised and it is, indeed, often used in practice. This is the "ratio" scale as opposed to the "natural" scale; it is not "unnatural," but is certainly less obvious and familiar. At first, therefore, care must be taken to understand what the scale is and how it is used.

The ratio scale is also called the logarithmic scale. This is because its construction is based on logarithms and some elementary knowledge of these is needed, as found in any ordinary text-book on arithmetic or algebra. Whereas a natural scale will space out equally any series of numbers which increase by a fixed amount, the object of a ratio scale is to

space out equally numbers which increase by a fixed multiple or percentage. Such a series is the following which increases ten-fold from each number to the next:

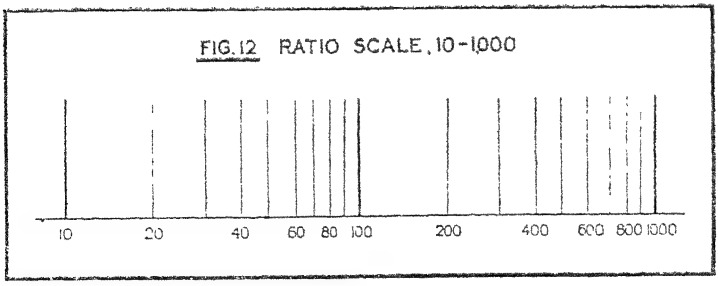
10; 100; 1,000; 10,000; 100,000; . . .

The logarithms of the numbers in such a series increase by fixed amounts; they are the series of integers (1, 2, 3, 4, 5, . . .) in the example quoted. This is the clue to the ratio scale. If we mark off numbers on a scale by taking distances along the scale as logarithms, then we have a scale with the required property.

A ratio scale can be constructed on ordinary squared paper with the aid only of tables of logarithms. Suppose we wish to have a scale ranging from 10 to 1,000. Then:

No.	Log.	No.	Log.	No.	Log.	No.	Log.
10	1·000	60	1·778	100	2·000	600	2·778
20	1·301	70	1·845	200	2·301	700	2·845
30	1·477	80	1·903	300	2·477	800	2·903
40	1·602	90	1·954	400	2·602	900	2·954
50	1·699			500	2·699	1,000	3·000

Use the *logarithms* to measure off along the chosen axis of the scale and mark the *numbers* at the points obtained. This is shown in Fig. 12.



The scale can be tested to verify that numbers in the same ratio are represented at equi-distant points. There is a rise of 50 per cent from 60 to 90 and from 400 to 600. On the ratio scale the distance between the first two points equals that between the second two. Equal distances on a ratio scale

always show equal percentage changes. This is no more than a reflection of the properties of logarithms; the logarithm of a product is the sum of the separate logarithms and of a ratio the difference of the separate logarithms. With logarithms a ratio turns into a difference.

3.9 *Graphs on Ratio Scales.* Any time series can now be plotted in two ways. Time is measured along the horizontal axis on a natural scale; the variable is measured along the vertical axis *either* on a natural *or* on a ratio scale. A graph of the second kind is the new construction; it is often called a *semi-logarithmic graph* since the ratio or logarithmic scale is used on one of the two axes of the graph.

Bank investments in the period 1921–45 (data of Appendix I,

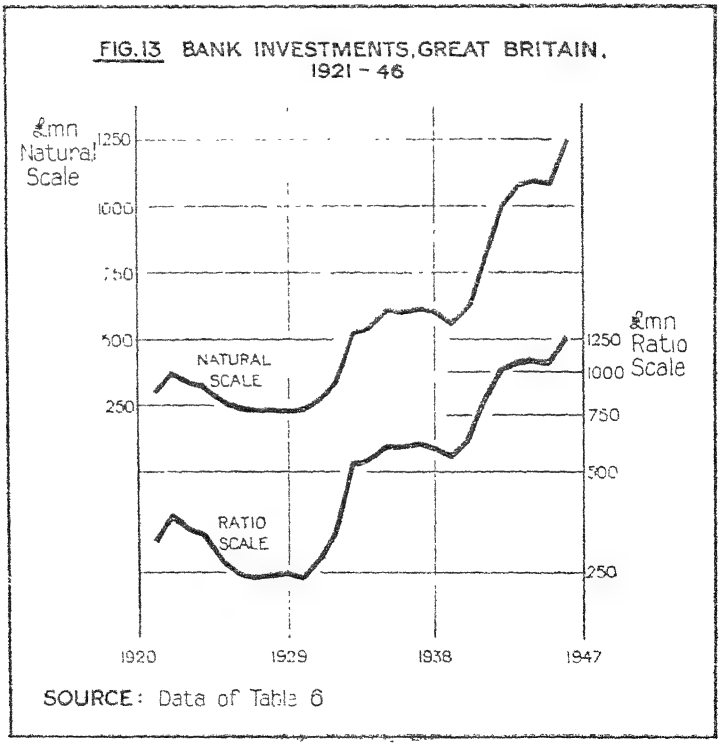


FIG. 13 BANK INVESTMENTS, GREAT BRITAIN, 1921 – 46

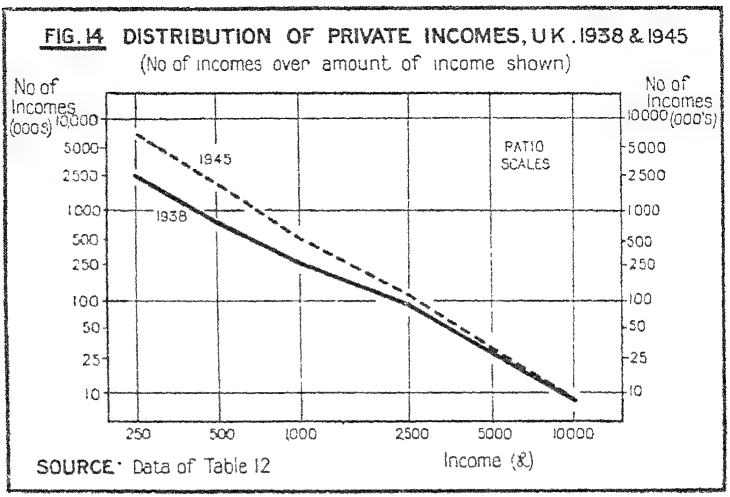
SOURCE: Data of Table 6

Table 6) are shown graphically in Fig. 13 on natural and ratio scales. An inspection of this diagram will show how the ratio scale works in practice. Consider the increase in investments from £309 mn. to £372 mn. in 1921–22 and that from £837 mn. to £1,006 mn. in 1941–2. On the natural scale the second rise (of £169 mn.) is shown as between two and a half and three times the first rise (of £63 mn.). On the ratio scale, however, the two rises are about equal since each is a little over 20 per cent. Which is the preferable representation? This depends on the point of view, on what the graph is to show. It may be that we regard equal percentage changes in investments as of equal significance, e.g., the change in 1941–2 as of the same importance as that in 1921–2. If this is so, then the ratio scale is the appropriate one.

We use a ratio scale, therefore, when we wish to show proportional or percentage changes in the variable and not absolute changes. The ratio scale, in particular, is useful when the variable shows a great range of variation, as during a war. It damps down the amplitude of the variations on high values of the variable and enables us to look at them in proportion.

It should be noticed that there is no zero base line on a graph drawn with a ratio scale. The zero point of the scale is always off the bottom of the paper (since the logarithm of zero is minus infinity). This is often a positive advantage. Two or more time series can be drawn on one graph on a ratio scale without the difficulty met with the natural scale (3.3 above). The same ratio scale must be used for each series so that a given percentage increase appears as the same jump on one as on other series. But the scale itself can be moved up or down without affecting the comparison. This is seen by an inspection of Fig. 13.

Ratio scales can be used for other diagrams than time graphs. Moreover, though the horizontal (time) axis on a time graph is always taken on a natural scale, in other diagrams it is sometimes appropriate to take ratio scales along both axes. In such a case, a *logarithmic graph* is drawn and proportional changes in one variable are related to proportional changes in a second. A good example is the distribution of incomes shown in Table 12 of Appendix I. Because of the great range both of incomes and of numbers with various incomes, it is inappro-



priate, indeed practically impossible, to represent these data on natural scales. If we concentrate on proportional changes, we can make a logarithmic plotting of the data as in Fig. 14. The distribution is first cumulated but in the direction opposite to that used in 3.6; we write the number of incomes £250 and over, the number £500 and over, and so on:

Income Range	No. of incomes (000's)	
	1938	1945
£250 and over	2,545	7,592
£500 and over	800	2,192
£1,000 and over	300	542
£2,000 and over	105	132
£10,000 and over	8	8

A cumulative graph showing the distribution each year is then drawn with a ratio scale both for income and for numbers. The points of the cumulative graph each year are seen to lie close to a straight line. This shows up a relation (known as Pareto's law) which is often found to hold, approximately, between the number of incomes and income size. The number

of incomes over a certain level falls proportionately as the income level rises proportionately. The logarithmic graph provides at least a rough method of seeing whether this law is appropriate to actual data.

CHAPTER IV

DERIVED STATISTICS

4.1 *Analysis of Statistical Tables.* Graphs and diagrams help to show up trends and relations but they do not define or measure them precisely. This can be achieved only by calculations on the numerical data and, in particular, by the derivation of figures to summarize and relate the significant facts in a table. The main purpose of statistical analysis is to make comparisons. A single figure has no meaning by itself; it only becomes significant and "alive" when compared explicitly or implicitly with another figure. Our first task in analysis is to make the comparisons explicit, to express the relation between one figure and another.

A simple example will serve to drive home this point. The numbers employed in building in Great Britain are given as 1,184,000 in mid-1946 (Appendix I, Table 2). Is this large or small? Is the building trade expanding or declining? The figure itself gives no answer to these and other relevant questions. To the expert, perhaps, the number means something but only because he knows some other fact to which he automatically relates this one—say the numbers employed in building in an earlier year. We must make such a comparison explicit. Table 2 also gives the numbers employed in building in mid-1945, namely 722,000. Hence, during a year of reconversion from war, building employment increased considerably. The figures in col. (5) of Table 2 are written to bring out this fact. The entry for building shows that employment in this trade in mid-1946 was

$$\frac{1{,}184{,}000}{722{,}000} \times 100 = 164 \text{ per cent}$$

of that in mid-1945, a rise of 64 per cent in a year. Col. (5) does not add anything to what is shown in the other columns of the table, but it does make precise and explicit the relations between the figures of the table.

4.2 *Ratios and their Specification.* The simplest derived statistics, as in the last example, are obtained by dividing one figure by another to give a ratio. The ratio is often, but not always, expressed conveniently as a percentage. Ratios can be classified into three broad types. Firstly, the part can be related to the whole. Examples are the percentage of the population in a definite age group (Appendix I, Table 11), and the percentage of workers employed, i.e., 100 times the ratio of those in employment to the total of all workers, employed and unemployed (Appendix I, Table 6). If the calculation is made for each of the parts making up the whole, then a series of ratios is obtained adding up to unity or 100 per cent. This is illustrated in Table 11 where the age distributions are expressed in percentage form so that they can be compared directly.

Secondly, one figure can be related to a similar figure, e.g., at a different time or place. A comparison over time, one of the commonest uses of ratios, is illustrated in Table 2 of Appendix I, where employment at one date is expressed in col. (4) or (5) as a percentage of that at an earlier date. Any one comparison of this kind can be made either way. For example, in Table 2, building employment in mid-1946 (1,184,000) is 164 per cent of the corresponding employment in mid-1945 (722,000). With the comparison reversed, employment in mid-1945 is 61 per cent of that at the later date. The relation between these two expressions of the same thing is one of multiplication of ratios: $1{\cdot}64 \times 0{\cdot}61 = 1{\cdot}00$. One ratio is the reciprocal of the other. It is *not* a relation of addition as might easily be supposed; an increase of 64 per cent from 1945 to 1946 is *not* the same as a decrease of 64 per cent backwards from 1946 to 1945.

The process of relating figures at different times can be extended to a whole series of dates; one of the dates is selected as the basis for comparison and all the others compared with it in succession. So, in Table 5 of Appendix I, earnings in each year (1880–1938) are expressed as a percentage of earnings in one particular year (1914). This is the basis of the method of index numbers (Chapter VI).

Finally, one figure can be related to another of different but associated type. The death rates of Appendix I, Table 11,

provide a simple example; here the number of deaths in an area *in a given period* is related to the population of the area *at a given date*. The derived figure is, in this instance, expressed as a rate per 1,000. There are many other examples of varying types—the number of persons per family obtained by dividing the total of persons by the number of families; rent per acre by division of rent paid by number of acres occupied; imports of tea (in lb. or £'s) per head of the population derived from total imports and total population.

There is little difficulty with ratios of the first two types beyond the need for ensuring that the figures related do in fact correspond. The units of measurement of the two figures are the same and the ratio is a simple figure. It suffices, for example, to say that, in England and Wales in 1938, those aged seventy-five and over were 2·5 per cent of the total population. There is some flexibility in the mode of expression of the ratio. Strictly, the ratio is a figure obtained by direct division, and it is less than unity when a part is related to the whole. It is often convenient to multiply the ratio by 100 and write it as a percentage. There is nothing special about a percentage; it is simply a device, generally accepted, for avoiding a ratio starting with a decimal point and zeros. So, 2·5 per cent is just another expression of the ratio 0·025. Occasionally, the ratio is expressed in other ways, e.g., per 1,000 or per 1,000,000.

There are more difficulties in devising and expressing ratios of the third type, which are like many concepts (e.g., velocity, acceleration) used in mechanics and other natural sciences. The figures related are dissimilar and often expressed in unlike units or for different periods. The ratio needs careful specification as so much of something per so much of something else. Often the time period involved requires explicit reference by the addition of per such and such a period to the specification. For example, a death rate is a number per 1,000 of the population. But this is not quite good enough, since the deaths occur over a period (month, quarter, year) while the population is given at a given date. A complete expression of the rates of Appendix 1, Table 11, would read the number of deaths per year per 1,000 of the mid-year population. Again, rent per acre needs to be "dressed-up" in

a similar way, e.g., as annual rent in shillings per acre occupied at a date.

One of the big problems in handling such ratios in practice is to obtain *appropriate* figures for comparison. One figure is to be related to another as a standard of reference, but it must be a suitable standard. Often a completely appropriate comparison cannot be devised and an approximate relation must suffice. Two examples will illustrate the difficulties. The expression of imports of tea per head of the population of this country is a fair but not a perfect comparison since, though most members of the population drink tea, some take only milk or coffee. Again, a marriage rate is obtained by dividing the number of persons married by some total. What total? Clearly not the total population which includes many not "at risk" as regards marriage—young children, those already married and others. A good ratio, still only approximately correct, is to the number of single, widowed and divorced persons aged fifteen and over.

4.3 *Approximations.* All statistical data are subject to errors in collection (see 1.4). It follows that any figure quoted is an approximation to the correct but unknown value. According to official returns for 1938, U.K. exports of beer were valued at £1,142,896, and exports of fruit juice and table waters at £223,490. These figures are based, not on actual sailings of ships with cargo, but on exporters' returns received at the Board of Trade in the year. Even with this restriction they do not pretend to be accurate to the last £ since there are always errors which escape correction in quantities exported and in their valuation. In fact, the figures are approximations and we do not know how much they may err.

Assessing the methods of collection, we may decide that the valuation may be as much as £1,000 out either way. We can then write exports of beer as £1,142,896 ± £1,000, the correct value lying between £1,141,896 and £1,143,896. There is clearly no point in carrying the last three digits and it is better to write the value as £1,143,000 ± £1,000 with a range from £1,142,000 to £1,144,000. This points to the need for rounding such approximate figures.

There is a further reason for rounding statistical data. A

figure running into millions of £'s, for example, is not easily appreciated if all the seven or more digits are shown. It becomes more comprehensible if rounded to so many millions or hundreds of thousands of £'s. Even if the figure were completely accurate, it would be unnecessary to carry it in full in most statistical presentations and calculations. It is a waste of time to calculate with more precision than we need.

It is common practice to round statistical figures. The need for rounding usually arises because the basic figures are approximations. But it is reinforced by the practical convenience of having rounded figures in statistical computations and for presentation of results.

4.4 *Rounding Statistical Figures.* The general process of rounding is to take each recorded value to the *nearest* digit considered. There are two methods to distinguish and, we shall see, each has its appropriate uses. The first method is to round to a specified unit, e.g., to the nearest 1,000. All digits after the last considered are ignored, except that they determine whether the last digit is left or raised by one. Thus, in rounding export values to the nearest £1,000, we write £223,490 as £223,000, but we write £1,142,896 as £1,143,000.

The second method is to round to so many significant figures. The units or the place of the decimal point are now ignored, and only the first so many digits from the left are written down, the last being taken to the nearest digit and followed by zeros. Thus, in rounding export values to four significant figures, we write £223,490 as £223,500, and £1,142,896 as £1,143,000.

The various methods and degrees of rounding, and the appropriate notations, are illustrated in the table:

Approximation	Exports of beer		Exports of fruit juice, etc.	
	£	£000	£	£000
Original	1,142,896		223,490	
Nearest £100	1,142,900	1,142·9	223,500	223·5
Nearest £1,000	1,143,000	1,143	223,000	223
Nearest £10,000	1,140,000	1,140	220,000	220
5 significant figures	1,142,900	1,142·9	223,490	223·49
4 significant figures	1,143,000	1,143	223,500	223·5

Zeros are written after the digit rounded up or down. It is often convenient to drop some or all of the terminal zeros, making use of a decimal place if necessary, as illustrated in the column headed £000. So, 1,142·9 in £000 is an alternative way of writing £1,142,900, each being rounded to the nearest £100; the former is simply more convenient for many purposes.

The need for rounding is imperative in calculating derived statistics. The division of one number by another can generally be continued to an indefinite number of decimal places; the question is where to stop. The answer is partly a matter of convenience and partly dependent on the accuracy of the figures used in the ratio (see 4.6). For example, in Appendix I, Table 2, col. (5), employment at mid-1946 is expressed in percentage of that at mid-1945 and each percentage rounded to one decimal place. Thus, for building,

$$\frac{1,184}{722} \times 100 = 163{\cdot}989 \ldots = 164{\cdot}0 \text{ per cent.}$$

Again, in col. (5) of Appendix I, Table 3, average export values in 1938 are computed by dividing values by quantities and carried to three or four significant figures. So

$$\frac{1,142,896}{271,094} = 4{\cdot}2158 \ldots = 4{\cdot}22 \text{ to 3 significant figures}$$

is the average value (£ per barrel) of exports of beer.

One point of practical importance can be noticed in passing. When we write 164·0 per cent, as above, we mean that the figure is rounded to one decimal place. The true figure is 164·0 ± 0·05, between 163·95 and 164·05. The zero after the decimal point is significant and must be written; 164·0 is not the same as 164, for the latter means 164 ± 0·5 or between 163·5 and 164·5.

4.5 *Errors in Sums and Differences.* We shall now define the error in a rounded figure in a particular sense. The (unknown) correct figure lies in a certain range, so much above and so

much below the rounded figure. We take the *error* as the greatest variation up and down so that the correct figure is in the range: rounded figure $\pm$ error. For example, per 100 of the 1938 population, there were 14·6 aged 5 and under 15, and 15·9 aged 15 and under 25 (Appendix I, Table 11). These figures should be written 14·6 $\pm$ 0·05 and 15·9 $\pm$ 0·05 with an error of 0·05 each way.

Suppose we add two rounded figures. In our example, the number aged 5 and under 25 is 14·6 + 15·9 = 30·5 per 100 of the population, from the rounded figures. The correct sum may be as low as 14·55 + 15·85 = 30·4 or as high as 14·65 + 15·95 = 30·6. Hence the sum is 30·5 $\pm$ 0·1 with an error of 0·1 each way. The error in the sum is the sum of the separate errors (0·05 each) in the original figures.

A similar result is obtained in subtraction. Per 100 of the population, the number of those aged 15 and under 25 exceeds the number aged 5 and under 15 by 15·9 — 14·6 = 1·3, from the rounded figures. The correct difference may be as low as 15·85 — 14·65 = 1·2 or as high as 15·95 — 14·55 = 1·4. The difference, in fact, is 1·3 $\pm$ 0·1. The error is again the sum of the two original errors of 0·05 each. These are perfectly general results which can be extended:

The error in a sum or difference of any number of rounded figures is the sum of the errors in the separate figures.

In adding up a column of rounded figures, there is no point in mixing items of different accuracy; each figure should be rounded off to the same unit. The error in the sum is then a multiple of the error in each figure, the multiple being the number of figures added. In col. (8) of Table 3, Appendix I, 1938 exports are valued at 1935 prices and each of the six values is rounded to £1,000. The sum (in £000) of the column is 13,378 with an error of 6 $\times$ 0·5. It can be written 13,378 $\pm$ 3, and the correct sum lies between 13,375 and 13,381.

The same problem appears in a different guise when the separate items *and* the sum are given and all are rounded. For example, col. (10) of Table 3 is obtained as follows:

Commodity	Original values (£)	Rounded values		
		to £1,000 (£000)	to £10,000 (£000)	to £100,000 (£ millions)
Spirits	11,361,617	11,362	11,360	11·4
Beer	1,142,896	1,143	1,140	1·1
Fruit juice and table waters	223,490	223	220	0·2
Cocoa preparations:				
With sugar	547,527	548	550	0·5
Without sugar	264,771	265	260	0·3
All other items	283,561	284	280	0·3
Total	13,823,862	13,825	13,810	13·8

One of the rounded columns adds as it stands; the other two do not. This is to be expected since the sum of six rounded figures may be in error by anything up to 3 either way in the last digit. So, in the second column, the sum of the items is $13,825 \pm 3$; the actual total is rounded to 13,824, within the range.

This point always arises in writing a column of percentages to an actual total of 100 per cent. The separate percentages, each being rounded, need not add exactly to 100. In Table 11 of Appendix I, col. (3) or (4) contains nine percentages to one decimal place, and the sum is subject to an error of $9 \times 0{\cdot}05$, i.e., it lies in the range 99·55 to 100·45. Actually, col. (3) adds to 99·9 and col. (4) to 100·0.

We must not be afraid to show a column of rounded figures which do not add exactly to the entry at the bottom. They must be rounded correctly and not "doctored." If explanation is needed, a note can be added: "figures do not necessarily add to totals because of rounding."

4.6 *Errors in Products and Quotients.* In sums or differences, each figure is to be rounded to the same unit, e.g., to one decimal place. In products or quotients, however, the relevant fact is the number of significant figures used. For example, the accuracy of $1{\cdot}2 \times 0{\cdot}76 = 0{\cdot}912$ does not depend on the number of decimal places in the figures multiplied, but on the fact that each is given to two significant figures. The product

can be found from $12 \times 76 = 912$ and the decimal point inserted in the end.

The useful concept is now that of *relative error*. If 1·2 is rounded to two significant figures, the error is 0·05 and the relative error is $\frac{0{\cdot}05}{1{\cdot}2} \times 100 = 4{\cdot}17$ per cent. Similarly the relative error in 0·76 to two significant figures is $\frac{0{\cdot}005}{0{\cdot}76} \times 100 = 0{\cdot}66$ per cent. The smallest value of the product of 1·2 and 0·76 is $1{\cdot}15 \times 0{\cdot}755 = 0{\cdot}868$, and the largest value is $1{\cdot}25 \times 0{\cdot}765 = 0{\cdot}956$. Hence, the product is $0{\cdot}912 \pm 0{\cdot}044$ with a relative error of $\frac{0{\cdot}044}{0{\cdot}912} \times 100 = 4{\cdot}83$ per cent. Now, 4·83 happens to be equal to the sum of 4·17 and 0·66; the relative error in the product is the sum of the relative errors in the two numbers multiplied. This is no accident. It is a general, though approximate, rule which applies equally to multiplication and division:

> *The relative error in a product or quotient of two rounded figures is approximately the sum of the relative errors in the separate figures.*

Errors are additive, and so increased, in the process of multiplication or division. The result is correct to a number of significant figures no greater (and usually less) than either of the original figures. In our example, the product of 1·2 and 0·76 is in the range from 0·868 to 0·956. It can be quoted only as 0·9 to one significant figure, as compared with two significant figures in the numbers multiplied. (Even this is not quite safe since it is just possible that the product exceeds 0·95.) This suggests the following rough rule for use in practice:

> *It is generally safe to write a product or quotient as correct to one less significant figure than the less accurate of the two values in the product or quotient.*

The problem also arises in converse form—how many figures should be carried in the original numbers to give products and quotients to a specified degree of accuracy? It

is as well to have a margin in deciding this point. A rule which is generally quite safe is to include one more significant figure in the computation than is needed in the answer. If a ratio is required to two significant figures, for example, it is usually safe to work through the calculation with three significant figures and to drop the last digit in the result.

4.7 *Some Examples and Warnings.* In a reply to a parliamentary question (Hansard, 14th November, 1945), the approximate number of Moslems in the Empire was given as follows:

India	92,000,000
Dominions	161,750
Colonies	13,325,000
Total	105,486,750

The first figure is rounded to the nearest million. It is not only a waste of time to show the last six digits in the sum; it is positively misleading. The sum should be given as 105 millions approximately.

The following is derived from Table 3 of Appendix I:

Exports in 1938	Value (£000)	Error	% Error	Signif. figures
Fruit juice, etc.	223	$\pm$ 0·5	0·224	3
Total beverages	13,824	$\pm$ 0·5	0·004	5

Exports of fruit juice, etc., are to be written as x per cent in value of total exports of beverages. So:

$$x = \frac{223}{13{,}824} \times 100 = 1{\cdot}613 \ldots$$

The relative error in x is (approximately) the sum of the separate relative errors, or 0·228 per cent. The actual error is then

$$1{\cdot}613 \times \frac{0{\cdot}228}{100} = 0{\cdot}004.$$

Hence $$x = 1{\cdot}613 \pm 0{\cdot}004$$

and we can write $x = 1{\cdot}6$ for certain and $x = 1{\cdot}61$ with fair safety. This also follows from the rough rule. The less accurate

of the numbers divided is correct to three significant figures, and so x can be quoted as 1·6 to two significant figures.

From Appendix I, Table 11, the percentage distribution of the population of S.W. England in 1938 is:

Age (years)	Population %	Population (000's) Reconstructed	Actual
0–	20	417	410
15–	30	625	625
35–	27	562	564
55–	23	479	484
Total	100	2,083	2,083

The percentages are rounded to the nearest whole number, i.e., to two significant figures. Suppose we know only these rounded percentages and the total population (2,083,000). Then the distribution of the population (in 000's) may be reconstructed by multiplying each percentage by $\frac{2,083}{100}$ as shown above. But, since the percentages are to two significant figures, the reconstructed numbers can only be relied upon to one significant figure. In this instance, we can check by comparing with the actual numbers of the population to three significant figures. In general, when we reconstruct unknown numbers from rounded percentages we must remember that we have no more accuracy than in the percentages themselves —that we may be getting very rough approximations indeed. We sacrifice accuracy in writing the percentages and we cannot recover it by multiplication back again.

Other types of reconstruction can be even more risky. The first two columns below show data on the industrial population (excluding government and the Forces) of Great Britain:

	June, 1939 Numbers (000's)	June, 1946	
		% of June, 1939	Est. numbers (000's)
Employed	16,535	93	15,378
Not employed	1,270	...	...
Total	17,805	92	16,381

We have to deduce the change in the numbers not employed. The last column is obtained by applying the given percentages to the numbers in 1939. This apparently turns the trick. The number not employed in 1946 is derived by difference: 16,381 — 15,378 = 1,003 in 000's. This is 79 per cent of the 1939 figure.

Actually, from Table 2 of Appendix I (from which these data are drawn), the number not employed in 1946 is 1,076 in 000's or 85 per cent of 1939. Our calculation is very far out. Why? The reason is that we have ignored the errors in the figures. The 1939 numbers are to the nearest 5,000, and the percentage changes are to 1 per cent. The number employed in 1946 lies between $16{,}532\frac{1}{2} \times 92\frac{1}{2} = 15{,}293$ and $16{,}537\frac{1}{2} \times 93\frac{1}{2} = 15{,}463$. It should be written $15{,}378 \pm 85$. Similarly, the total number in 1946 is $16{,}381 \pm 91$. (These can be derived also by use of the rule in 4·6 above.) Hence, the difference has an error which is the sum: $85 + 91 = 176$. The result of our calculation is $1{,}003 \pm 176$, which clearly shows the inaccuracy of the whole process.

The difficulty here is that the figure sought is a small difference between two large and nearly equal figures. All these figures have errors, of the same general size, fairly small relative to the original numbers (of employed and total) but quite large relative to the difference (number not employed). In short, it is very risky to estimate a small difference between two nearly equal figures subject to error.

4.8 *Biassed and Unbiassed Errors.* So far we have concentrated upon the worst error that can occur. The largest error in adding two figures is the sum of the separate errors, but this happens only when there is a maximum error in one figure *and* a maximum error in the same direction in the other. In practice, we expect something better, some cancelling of errors.

Errors in figures are *unbiassed* if those in one direction are as likely to occur as those in the other direction. This is usual when figures are rounded. Errors now tend to cancel out when the figures are added or multiplied. For example, the sum of twenty percentages rounded to one decimal place *can* vary from 99·0 to 101·0, but it is not *likely* to differ from

100·0 by more than 0·2 or 0·3. On the other hand, unbiassed errors are not so helpful when figures are subtracted or divided; they do not necessarily tend to cancel.

If errors tend to occur in the same direction, then they are *biassed*. They will then tend to cumulate in addition or multiplication. But something may be gained in a difference or ratio; the difference (or ratio) of two figures with biassed errors may be more accurate than the figures themselves. Biassed errors occur quite often in practice. For example, in obtaining the total value of retail sales from returns from individual stores, we may have a few returns missing each month so that the total is biassed downwards. The difference (or percentage change) in sales from one month to another can be more reliable than the monthly totals.

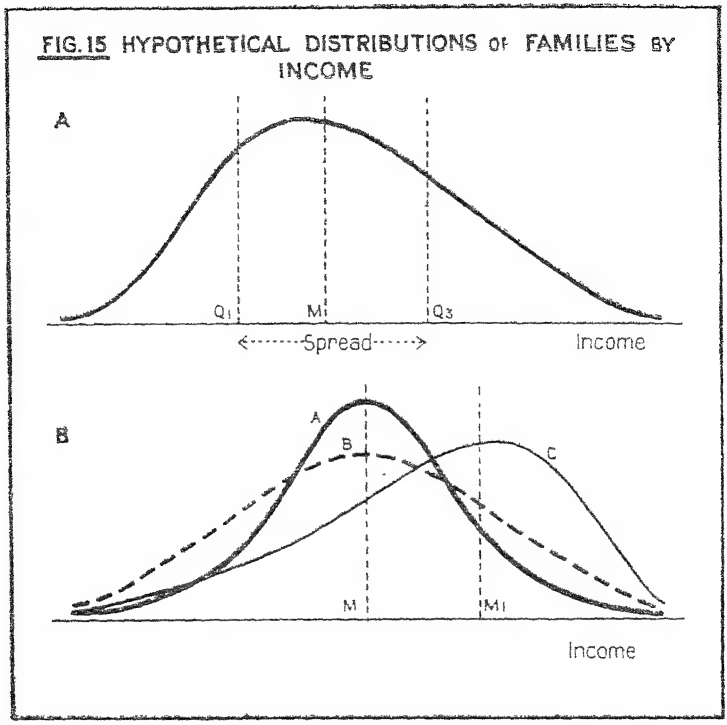
CHAPTER V

FREQUENCY DISTRIBUTIONS

5.1 *Summarization of Frequency Distributions.* The value of a variable character is specified for each of a given set of items, e.g., the income of each family in a given group or the age of each individual in a given population. Such data can be summarized into a frequency distribution (1.8 above) and can be represented graphically as a block diagram (3.5 above). Here the whole range of the variable is divided into a convenient number of classes and the number of items in each class is obtained. It is sometimes convenient to imagine an ideal or theoretical frequency distribution in which the number of classes increases and gets indefinitely large. The block diagram includes more and more rectangles, each thinner and thinner. In the limit, the diagram can be pictured as a smooth curve and the area under the curve (between any pair of vertical lines) represents the number of items occurring between specified values of the variable.

A hypothetical frequency curve of families distributed by income is shown in Fig. 15A. This is the kind of shape often taken by frequency distributions in practice. Other forms do occur, if rather less frequently. Sometimes the curve has two or more peaks instead of one. This can happen particularly when the data are not homogeneous, as in the distribution of heights of a group consisting of some young children and some adults. Sometimes the curve does not have a peak at all, but takes the form of a J or inverted J (as in Fig. 10), or even the form of a U.

Though more manageable than the basic data, a frequency distribution remains a complex expression of the material. Further summarization is needed and one way is to pick out some representative figures to describe the distribution. The first of such figures is an *average* or central representative value of the variable, about which all the other values are grouped. The average income M in Fig. 15A corresponds to a vertical



line splitting the area under the frequency curve into two roughly equal parts. A second figure is needed to supplement the average in indicating whether the values of the variable are concentrated about the average or widely dispersed. This is a measure of the spread or *dispersion* of the distribution. Fig. 15A shows one possible measure, the spread between two vertical lines which, with the vertical line at the average, divide the area of the frequency curve into four roughly equal parts. Further figures can be defined, for example, to describe *skewness* by indicating whether the distribution is symmetrical about the average or is "humped" to one side or the other.

Some of the possibilities are illustrated in Fig. 15B. The two symmetrical curves A and B have the same average M and differ only in that the variable (income) is more widely

dispersed among families in B than in A. The curve C has a higher average income, M; it is fairly widely dispersed, more so than A certainly; unlike A and B, it is skew with a cumulation of families with income above the average and with a long "tail" corresponding to a few families with low incomes. It is clear from these examples that the concept of average is most useful but not sufficient in itself to summarize a distribution (e.g.) of families by income. An indication of dispersion is required to show that B has more families with high incomes, and more with low incomes than A. Measures of dispersion and skewness are needed to indicate that C has a higher average income and, at the same time, more very low incomes than A.

5.2 *Everyday Uses of Averages and Dispersion.* The concept of an average is generally appreciated. A batsman at cricket will make a great variety of scores in the course of a season. His achievements are summarized in the publication of his batting average, that is a score typical of the whole distribution of his scores in various innings. He may never make an actual score equal to his average; indeed the average is usually calculated as a number, such as 42·6, carried to one or more decimal places so that he could not possibly make such a score in one innings. Nonetheless, the average is accepted as representative.

The concept of dispersion is rarely as explicit as this, but it is often implicit in everyday usage. Two batsmen may have the same average of forty-five, but one may be regarded as a more reliable scorer than the other. The first batsman makes a good score of forty to fifty almost every time he goes in to bat; the other keeps up his average with a few very big scores in between many low ones. The distribution of scores is more dispersed (and also more skew) for the second than for the first batsman.

Such everyday examples can be multiplied indefinitely. When we say that house rentals are higher in Westminster than in West Ham, we have an average in mind, and we do not rule out the possibility that there are individual houses in West Ham for which more rent is paid than many in Westminster. Indeed, most prices used in practice are averages,

with or without an indication of dispersion. For example, stock market prices as quoted are averages of various transactions and a measure of spread (e.g., high and low quotations) may be given as well. Again, we may say that the inequality of incomes in the country is less now than some years ago; we mean that there has been a reduction in the dispersion, and perhaps also in the skewness, of the distribution of incomes amongst individuals. On the other hand, the fact that an average implies a dispersion about it is by no means always appreciated. For example, the statement has been made, with alarm, that no less than half the adults in this country have height below the average; this is a natural property of an average, not an indication of under-nourishment.

It remains to systematize, and to make precise, these everyday notions in framing specific definitions of average, dispersion and skewness. Such concepts can remain a little vague in ordinary speech, but they must be unambiguous in statistical work. The measurements in statistics must be precise and often complex; but the basic ideas are commonplace, here to obtain a few representative figures to bring out the main characteristics of a frequency distribution. In addition, the measures should be readily calculated and capable of easy manipulation in statistical analysis. As we shall see, there is not one, but several possible definitions we can adopt for (say) an average, all satisfying the basic requirement of a representative value of the variable. One measure may be more useful for this purpose and another for that. There is the same choice in measures of dispersion or skewness. In fact, we are to aim at precise and convenient, but not unique, definitions of these concepts.

5.3 *Median, Quartiles and Quartile Deviation.* The basic data of a frequency distribution always consist of a fully detailed list of values of the variable. It is convenient, in developing definitions and methods of calculation, to work first with one specific case of simple type. In this and the following sections, we take for illustration the distribution of the forty-five price relatives used in the *Statist* index number of wholesale prices in 1938 (average 1867-77 = 100). All the individual relatives, and two alternative methods of classifying

them into frequency distributions, are shown in Table 9, Appendix I.

As a first step, the original values can be rearranged to appear in ascending order of magnitude from the smallest to the largest (Table 9A). The *median* M of the distribution is defined as the value of the variable which comes halfway up this order. The *quartiles* are defined similarly as the values of the variable which come one-quarter and three-quarters of the way up the order. The lower quartile Q_1 is the smaller of the two, one-quarter the way up from the lowest value of the variable. The upper quartile Q_3 is the higher value, three-quarters the way up the order. Median and quartiles are all values of the variable and expressed in the units of the variable. In our example we obtain median and quartile price relatives. If the distribution were one of family income in shillings, the median and quartiles would also be family income in shillings.

We have forty-five price relatives in ascending order of magnitude and we are to pick out the relative at the half-way and quarter-way points. There is, however, a little ambiguity about these points which must be eliminated. If there were eleven items then Q_1, M and Q_3 would correspond to the 3rd, 6th and 9th respectively; there are two items before Q_1, two between Q_1 and M and so on. If there were twelve items, the best we can do is to fix Q_1 between the 3rd and 4th, M between the 6th and 7th and Q_3 between the 9th and 10th. There are then three items before Q_1, three between Q_1 and M and so on. At the risk of being rather arbitrary, we make a general rule:

For n items of a variable, all specified and arranged in ascending order of magnitude, Q_1 is the $\frac{(n+1)}{4}$ th, M the $\frac{(n+1)}{2}$ th and Q_3 the $\frac{3(n+1)}{4}$ th in the order.

In applying this rule, we take the nearest integer if an odd $\frac{1}{4}$ or $\frac{3}{4}$ is obtained and we take half-way between adjacent integers if an odd $\frac{1}{2}$ results. The conventional element in this

rule becomes less important the greater the number of items.

With forty-five items, Q_1 is located half-way between the 11th and 12th, M at the 23rd and Q_3 half-way between the 34th and 35th. From the order of price relatives (Table 9) we find:

$$Q_1 = 58 \qquad M = 86 \qquad Q_3 = 113$$

It is clear that the *median is a measure of average*. It is the point M in Fig. 15A which divides the area under the curve into two equal parts. There are as many items below the median value as there are above it. It is also evident that the *range from the lower to the upper quartile is a measure of dispersion*. This is the spread from Q_1 to Q_3 in Fig. 15A obtained by dividing the area under the curve into four equal parts. The middle 50 per cent of all items lie in this range of the variable; there are 25 per cent of the items outside the range at each end. The wider the range, i.e., the farther apart Q_1 and Q_3 are, the more dispersed is the distribution. The conventional measure of dispersion, however, is the *quartile deviation* defined as half the range:

$$\text{Quartile deviation } QD = \tfrac{1}{2}\,(Q_3 - Q_1)$$

In our example, the quartile deviation of the price relatives is $27\frac{1}{2}$. We can now say that the average (median) price relative is 86 with a dispersion (quartile deviation) of $27\frac{1}{2}$.

When the data are given, not in full, but as a grouped distribution, there is an inevitable loss of detail and the median and quartiles cannot be found accurately. They can only be approximated or estimated and a conventional assumption must be made before an estimate can be obtained. We can fix the class in which (say) the median falls, but we cannot say exactly where it lies within this class. Something must be assumed about the distribution of items in the class. A convenient *assumption* is that the items are *uniformly distributed* over the range of the class, the variable ranging uniformly from the lowest to the highest value of the class.

It then follows that we take the median the proportionate

way along the class in which it lies, but that we take $\frac{n}{2}$ rather than $\frac{n+1}{2}$ in fixing the median of n items.[1] In any particular case, we can verify that the same value of the median is then obtained whether we count up the distribution from the bottom or down it from the top. The quartiles are similarly located. The rule is:

For a frequency distribution of n items, given in classes, Q_1 is taken as the $\frac{n}{4}$th, M is the $\frac{n}{2}$th, and Q_3 is the $\frac{3n}{4}$th item from the bottom and estimated in the class in which it falls by applying a proportionate rule.

The estimation is made best with a cumulative table. For the first distribution of price relatives in 1938 (Table 93), the table cumulated from the bottom is:

	Under	Under	Under	Under	Under
Price relatives	24½	49½	74½	99½	124½
No. of items	2	5	16	29	36

This allows for the fact that price relatives are rounded to whole numbers. The median is to be located by the use of 22½ out of 45 items, in the class from 74½ to 99½. The class contains $29 - 16 = 13$ items, and the median is given by $22\frac{1}{2} - 16 = 6\frac{1}{2}$ out of these. Hence we take the median $6\frac{1}{2}$ thirteenths along the class from the lower end of 74½. Our estimate is:

[1]This can be checked by examining the assumed uniform distribution of items in the class containing the median. If there are r items in the class, divide the range of the class into r equal intervals and place one item at the centre of each. The median is the $\frac{(n+1)}{2}$th item but the use of $\frac{(n+1)}{2}$ and the proportionate rule fixes the *upper end* of the interval, when we count up the distribution from the bottom. The *centre* of the interval, where the median is put, is half an interval back, i.e. $\frac{n+1}{2} - \frac{1}{2} = \frac{n}{2}$ should be used to locate it.

$$M = 74\tfrac{1}{2} + \frac{22\tfrac{1}{2} - 16}{29 - 16}\,25 = 74\tfrac{1}{2} + \frac{6\tfrac{1}{2}}{13}\,25 = 87{\cdot}0$$

Similarly

$$Q_1 = 49\tfrac{1}{2} + \frac{11\tfrac{1}{4} - 5}{16 - 5}\,25 = 49\tfrac{1}{2} + \frac{6\tfrac{1}{4}}{11}\,25 = 63{\cdot}7$$

$$Q_3 = 99\tfrac{1}{2} + \frac{33\tfrac{3}{4} - 29}{36 - 29}\,25 = 99\tfrac{1}{2} + \frac{4\tfrac{3}{4}}{7}\,25 = 116{\cdot}5$$

The same table can be cumulated the other way:

Price relatives	Over $174\frac{1}{2}$	Over $149\frac{1}{2}$	Over $124\frac{1}{2}$	Over $99\frac{1}{2}$	Over $74\frac{1}{2}$	Over $49\frac{1}{2}$
No. of items	2	4	9	16	29	40

The estimation of the median and quartiles, reading down from the top, is:

$$M = 99\tfrac{1}{2} - \frac{22\tfrac{1}{2} - 16}{29 - 16}\,25 = 99\tfrac{1}{2} - \frac{6\tfrac{1}{2}}{13}\,25 = 87{\cdot}0$$

$$Q_1 = 74\tfrac{1}{2} - \frac{33\tfrac{3}{4} - 29}{40 - 29}\,25 = 74\tfrac{1}{2} - \frac{4\tfrac{3}{4}}{11}\,25 = 63{\cdot}7$$

$$Q_3 = 124\tfrac{1}{2} - \frac{11\tfrac{1}{4} - 9}{16 - 9}\,25 = 124\tfrac{1}{2} - \frac{2\tfrac{1}{4}}{7}\,25 = 116{\cdot}5$$

The same estimates are obtained.

A similar calculation for the second distribution (Table 9c) of the price relatives in 1938 gives:

$$Q_1 = 63{\cdot}4 \qquad M = 87{\cdot}0 \qquad Q_3 = 112{\cdot}0$$

A comparison of these estimates with the correct values from the full data shows that there are considerable differences arising from the way in which the data are grouped and from the assumption made about distribution of items within a class. These differences diminish, but never disappear entirely, when there are more items and more classes in the distribution.

5.4 *Arithmetic Mean and Standard Deviation.* A simple shorthand notation now becomes convenient. Write Σx for the sum of all figures in a column labelled x, the Greek letter sigma (Σ) standing for "sum of all figures like." Similarly,

Σx^2 can stand for the sum of the squares of the figures x, $\Sigma \log x$ for the sum of logarithms of the figures x and Σxy for the sum of all products formed by multiplying an item in the column x by the corresponding item in the column y.

The average most commonly used in everyday life is the "arithmetic" average obtained by adding up the values of the variable and then dividing by the number of the items. It is the type used, for example, in computing batting averages at cricket. The definition can be taken straight over into statistical analysis for the *arithmetic mean AM*:

$$AM = \frac{1}{n} \Sigma x$$

where x is the variable and n the number of items. In our example, the 45 price relatives add to the total $\Sigma x = 4{,}068$, and so:

$$AM = \frac{4{,}068}{45} = 90{\cdot}4$$

As a first step in defining a measure of dispersion to correspond to AM, we can subtract the value of AM calculated from each original value of the variable to obtain the deviation of the value from the mean. This will be negative or positive according as the value is below or above the mean. The forty-five deviations are:

− 71·4	− 37·4	− 33·4	− 20·4	− 9·4	2·6	14·6	26·6	48·6
− 66·4	− 36·4	− 31·4	− 13·4	− 7·4	2·6	15·6	40·6	70·6
− 55·4	− 36·4	− 30·4	− 11·4	− 4·4	4·6	18·6	42·6	80·6
− 54·4	− 35·4	− 23·4	− 9·4	− 2·4	7·6	19·6	45·6	85·6
− 52·4	− 34·4	− 20·4	− 9 4	1·6	11·6	25·6	46·6	94·6

The sum of these deviations is zero by definition. If we take account only of the size of the deviation, ignoring the sign, we can add, divide by the number of items and obtain the mean of the deviations. This is a measure of dispersion since it is clearly larger the more widely spread is the distribution. It is called the *mean deviation* and in our example, adding the deviations above with sign omitted,

$$\text{Mean deviation} = \frac{1{,}413{\cdot}2}{45} = 31{\cdot}4$$

The mean deviation is awkward to compute when the items are numerous and grouped into classes, and it is also not convenient for algebraic manipulation, because of the difficulty of distinguishing and dropping the sign of negative deviations. We can avoid this difficulty by squaring each deviation, obtaining a series of positive values. The mean of the squared deviations is also a measure of dispersion and it is called the *variance*. This is a highly important concept in more advanced work where it is possible to split the total variances into several parts each attributable to one of the factors causing variation in the original series. This method of "analysis of variance" is described in the technical literature.[1]

The main point we can notice here is that the variance is in units which are the square of the original units. For example, if family income in shillings is the original variable, the variance is in units of (shillings)^2. It is thus convenient, as a measure of dispersion, to take the square root of the variance in order to get a figure in the original units. This gives the *standard deviation* as a measure of dispersion corresponding to the arithmetic mean:

$$\text{Standard deviation } SD = \sqrt{\frac{1}{n}\Sigma(x - AM)^2}$$

In our example, squaring and adding the deviations shown above,

$$\Sigma(x - AM)^2 = 70{,}451{\cdot}8 \quad \text{Variance} = \frac{70{,}451{\cdot}8}{45} = 1{,}565{\cdot}6$$

and

$$\text{Standard deviation} = \sqrt{1{,}565{\cdot}6} = 39{\cdot}6$$

When the data are given grouped, with consequent loss of detail, the mean and standard deviation must be estimated from some assumption about the distribution of items within each class. The convenient *assumption* is that the items are *concentrated at the centre* of each class. This assumption is different from that adopted for the median and the estimates obtained must be expected to differ also. It is found, in fact, that AM is the same estimated on the assumption of concentration at the centre as on the assumption of uniform

[1] *See* 7.7 below and Tippett, Appendix II, Ref. (17), or R. A. Fisher, Appendix II, Ref. (13).

distribution, but that *SD* has different estimates on the two assumptions. The assumption now made happens to be more convenient for purposes of calculation and, in adopting it, we must remember that our results are never more than estimates.

For the first distribution of 1938 price relatives (Table 9B) the whole calculation can be set out as follows:

Range	No. of relatives	Centre of range	Product (1) × (2)	Deviation (2) − AM	Square of deviation	Product (1) × (5)
	(1)	(2)	(3)	(4)	(5)	(6)
0–24	2	12	24	− 79·4	6,304·4	12,609
25–49	3	37	111	− 54·4	2,959·4	8,878
50–74	11	62	682	− 29·4	864·4	9,508
75–99	13	87	1,131	− 4·4	19·4	252
100–124	7	112	784	20·6	424·4	2,971
125–149	5	137	685	45·6	2,079·4	10,397
150–174	2	162	324	70·6	4,984·4	9,969
175–200	2	187	374	95·6	9,139·4	18,279
Total	45		4,115			72,863

$$AM = \frac{4,115}{45} = 91{\cdot}4 \qquad SD = \sqrt{\frac{72,863}{45}} = 40{\cdot}2$$

Notice that the validity of col. (3), and the later columns, depends on the assumption of concentration at the centre of each class. The determination of the centres in col. (2) depends on the correct interpretation of the classes. For example, the range "25–49" is to be taken as 24·5 to 49·5, since each relative is rounded to the nearest whole number, and the centre is then 37.

A similar calculation for the second distribution, with the first class taken as 0–49 for convenience, gives the estimates:

$$AM = 90{\cdot}9 \qquad SD = 40{\cdot}7$$

The effect of grouping, and of the assumption of concentration at the centre of each class, can be seen by comparing these estimates with the correct values, $AM = 90{\cdot}4$ and $SD = 39{\cdot}6$, obtained from the full data.

5.5 *Geometric Mean.* Measures of average other than the median and arithmetic mean can be devised and several are in fairly common use. The mode and harmonic mean are two which need not concern us here. Of greater importance is the *geometric mean.* This is a variant of the arithmetic mean in which a central value is obtained by multiplying, instead of adding, the original n values of the variable, and then by extracting the nth root. So:

$$GM = \sqrt[n]{\text{Product of } x\text{'s}}$$

where x is the variable and n the number of items.

If the data are given in full detail, then the calculation of GM is a matter of arithmetic, but the arithmetic is very lengthy if the multiplication and root extraction is carried through by hand. Fortunately, the use of logarithms simplifies the process and, in fact, reduces the calculation of GM to that of an ordinary arithmetic mean. By the properties of logarithms

$$\log\ GM = \frac{1}{n}\Sigma \log x$$

i.e., the logarithm of GM is the arithmetic mean of the logarithms of the original values.

In our example, the logarithms of the forty-five relatives are:

1·2788	1·7243	1·7559	1·8451	1·9085	1·9685	2·0212	2·0682	2·1430
1·3802	1·7324	1·7709	1·8865	1·9191	1·9685	2·0253	2·1173	2·2068
1·5441	1·7324	1·7782	1·8976	1·9345	1·9777	2·0374	2·1239	2·2330
1·5563	1·7404	1·8261	1·9085	1·9445	1·9912	2·0414	2·1335	2·2455
1·5798	1·7482	1·8451	1·9085	1·9638	2·0086	2·0645	2·1367	2·2672

$$\text{Hence } \Sigma \log x = 85{\cdot}8891 \qquad \log GM = \frac{85{\cdot}8891}{45} = 1{\cdot}90865$$

and $$GM = 81{\cdot}0$$

This can be compared with the median (86) and the arithmetic mean (90·4).

It is possible to devise methods of estimating GM from data grouped in a frequency distribution. These need not detain us here, however, since the geometric mean is usually needed only for simple distributions given in full detail.

5.6 *Comparisons by Averages and Dispersion.* Various measures of average and dispersion have been defined and calculated for one particular distribution, that of the forty-five price relatives of the *Statist* index number in 1938 (Table 9). Similar calculations can be performed for the other distribution of Table 9, that of the forty-five price relatives in the later year (1945). The results can then be assembled to give a comparison of the two distributions of price relatives, this being the object of these as of all statistical concepts. The advantages and disadvantages of the various measures can also be explored. The results are:

	1938 price relatives			1945 price relatives		
	Full details	Distrib. B	Distrib. C	Full details	Distrib. B	Distrib. C[1]
Averages						
Median *M*	86	87·0	87·0	142	144·1	143·5
Arithmetic mean *AM*	90·4	91·4	90·9	163·7	158·4	164·5
Geometric mean *GM*	81·0	...	...	144·4	...	...
Dispersion						
Lower quartile	58	63·7	63·4	111	108·5	109·0
Upper quartile	113	116·5	112·0	204·5	203·25	204·9
Quartile deviation *QD*	27·5	26·4	24·3	46·75	47·4	47·95
% of median	32	30	28	33	33	33
Standard deviation *SD*	39·6	40·2	40·7	84·5	83·9	85·0
% of arith. mean	44	44	45	52	53	52

[1]First class taken as 0–49 and last 300–499.

The estimate of a given average or measure of dispersion obtained from one classification is different from that obtained from the other classification of the same distribution, and each is different from the correct values computed from the complete data. The differences are not very large even in this example of only forty-five items and they tend to be smaller for a larger number of items. There are, however, always some differences, and the way in which a distribution is classified will always affect the estimates of average and dispersion derived from grouped data.

There is a simple relation between *AM* and *GM*. We have found that *GM* is less than *AM* in each of our examples. This is a general property for any set of positive numbers not all equal. There is no such relation for the median which can be either greater or less than the arithmetic (or geometric) mean. *AM* and *GM* are calculated by using all the data whereas *M* is found from the values in the middle of the range of data.

Hence, the former must be more affected by the particular values at the two extremes. Specifically, the addition of a few large values will pull up the *AM* considerably, *GM* less so and *M* scarcely at all. For example:

Value of Variable	*M*	*AM*	*GM*
3, 4, 5, 9	4·5	5·25	4·8
3, 4, 5, 9, 29	5·0	10·0	6·9

Similarly, the addition of a few small values will pull down *GM* considerably, *AM* less so and *M* scarcely at all.

M and *AM* are both easily calculated and easily interpreted. In ordinary use the choice between them depends largely on whether full weight is to be given to the extremes or not. In the first case, *AM* would be used; in the second, *M* is more appropriate. For more advanced work, however, *AM* has the advantage since it is capable of algebraic manipulation, whereas *M* is not.

The choice between *AM* and *GM* raises different questions. *GM* is a variant of *AM* and both can be handled algebraically. *AM* is generally to be preferred since it is easier to interpret and to compute. *GM* has certain particular uses, for example, in the development of index numbers and in handling ratio changes.

The uses of *QD* and *SD* as measures of dispersion are similar. If *M* is chosen as the average, then *QD* is the natural selection, and *AM* is similarly paired with *SD*. For a distribution which is fairly regular and symmetrical, there is an approximate relation between the two measures of dispersion; in practice *QD* is found to be about two-thirds of *SD*.

All these measures of average and dispersion appear in the units of the original variable. Sometimes, however, dispersion is better measured for purposes of comparison as a percentage or coefficient independent of units. For example, if the average income (in shillings) of one group of families is double that of another group, we may say that there is comparable inequality of incomes if the dispersion (again in shillings) is also double. A simple expedient is to express a measure of dispersion as a percentage of the corresponding average, *QD* in percentage of *M* and *SD* in percentage of *AM* as shown above. The second of these is termed the *coefficient of variation.*

The table of results gives a comparison of the distribution of prices in 1938 with that in 1945. The general level of prices rose from 1938 to 1945 and the extent of the increase is indicated by one or other of the averages of the forty-five separate price relatives. In terms of the *median*, the level of prices rose from 86 to 142 (in per cent of 1867–77) or an increase of 56 per cent. The increase is 80 per cent in terms of the *arithmetic mean*, from 90·4 to 163·7. A few exceptional prices show a very large increase from 1938 to 1945, and this is reflected in the larger rise shown by the arithmetic mean.

The main question arising in a comparison of the spread of prices is whether the dispersion of price relatives is relatively greater in 1945 than in 1938. This is seen to be the case, particularly if full weight is given to the very large price increases by the use of the standard deviation as a measure of dispersion. The standard deviation of the 1945 price relatives is 52 per cent of the mean; the corresponding figure for 1938 is 44 per cent. This increase in the spread of prices may be a particular result of the war of 1939–45 or it may be offered in support of the hypothesis that prices tend to become more dispersed as the period of comparison is lengthened.

5.7 *Skewness.* The measures of average and dispersion already calculated serve also to indicate the existence of skewness in a frequency distribution. A symmetrical distribution (with no skewness) has a "tail" of values above the centre exactly matched by a "tail" below the centre, and the median coincides with the arithmetic mean. A practical test of approximate symmetry is that these two averages are close together, closeness being judged relative to a measure of dispersion. If $AM-M$ is small as a percentage of QD or SD, then the distribution is approximately symmetrical. There is *positive skewness* if the upper "tail" is elongated and the lower "humped" together. Since a few large values pulls up the arithmetic mean more than the median, we expect that AM exceeds M for a positively skew distribution. Similarly, we expect that AM is less than M for *negative skewness* when the lower "tail" is elongated. The practical test for skewness, then, is that $AM-M$ is of significant size relative to QD or SD, and the sign determines whether the skewness is positive or negative.

Another test can be developed from the median and quartiles. A study of Fig. 15A indicates that M is midway between Q_1 and Q_3 for a symmetrical distribution while positive skewness pulls Q_3 away from M and negative skewness pulls Q_1 away from M. Hence the distribution is positively skew, symmetrical or negatively skew according as the mid-point between the quartiles is above, coincides with or below the median.

These are practical tests. It is more difficult to define a precise and satisfactory measure of skewness. One possible measure is $AM-M$ as a percentage of SD. A second is the difference between the mid-point of the quartiles and the median as a percentage of the quartile deviation, i.e.,

$$\frac{\frac{1}{2}(Q_3+Q_1)-M}{\frac{1}{2}(Q_3-Q_1)}\,100=\frac{(Q_3-M)-(M-Q_1)}{(Q_3-M)+(M-Q_1)}\,100$$

From the calculations (with complete data) of 5·6, the *Statist* price relatives in 1945 (Table 9) give + 26 per cent for the first measure of skewness and + 34 per cent for the second. The measures, however, are found to be inadequate for the 1938 price relatives, one giving a small positive and the other a small negative value. Hence, the distribution of price relatives has no definite skewness in 1938 but becomes markedly and positively skew in 1945.

A more precise measure, generally used in more advanced statistical methods, can be mentioned in passing. This measure extends the concepts of arithmetic mean and standard deviation. By analogy with the latter, the expression

$$\sqrt[3]{\frac{1}{n}\Sigma(x-AM)^3}$$

can be written and associated with skewness. For example, in a distribution with marked positive skewness, the large deviations above the mean, when cubed and magnified, more than counterbalance the smaller deviations below the mean and give a large positive value to the expression.

5.8 *Short Method of Calculating Mean and Standard Deviation.* The population of S.W. England has a different age composition from that of England and Wales as a whole, as shown in Appendix I, Table 11, for the year 1938. This

can be seen from a calculation of the mean age and the standard deviation of the age distribution of each population. Since the data are given in broad classes of age, only estimates of these measures can be obtained on the assumption that the numbers in each age class are concentrated at the mid-age of the class. These estimates can be found by the method already described (5·4), but exactly the same results can be obtained by a shorter calculation shown in the following table for the population of England and Wales:

Age in years	Centre of class		Population (000's)	Product	
	years	origin: 30 yrs. unit: 10 yrs.		(3) × (4)	(3) × (5)
(1)	(2)	(3)	(4)	(5)	(6)
0–	$2\frac{1}{2}$	$-2\frac{3}{4}$	2,818	$-7{,}749\frac{1}{2}$	21,311
5–	10	-2	6,025	$-12{,}050$	24,100
15–	20	-1	6,572	$-6{,}572$	6,572
25–	30	0	6,817	—	—
35–	40	1	6,034	6,034	6,034
45–	50	2	5,134	10,268	20,536
55–	60	3	4,240	12,720	38,160
65–	70	4	2,561	10,244	40,976
75–	$82\frac{1}{2}$[1]	$5\frac{1}{4}$	1,014	$5{,}323\frac{1}{2}$	27,948
Total			41,215	18,218	185,637

[1]Assuming class is effectively "75 and under 90."

Col. (3) is a modified form of col. (2) with age measured from a convenient but arbitrary point (30 years), and in terms of a convenient but arbitrary unit (10 years). All computations are made in terms of col. (3) and correction is made at the end to restore the original units (years of age). Notice, also, that col. (6) is col. (4) times the square of col. (3), but more easily written as col. (3) times col. (5).

This calculation shows how to deal with two minor complications. The classes of the distribution are not all of the same length; the first is shorter and the last longer. The centre of such classes needs to be located with care and the regular progression of the figures in col. (3) is disturbed. The last class is "open," i.e., 75 and over with no upper limit specified. A further assumption is needed for purposes

of calculation, here that the class is "75 and under 90" with centre $82\frac{1}{2}$.

The arithmetic mean is now derived from the total of col. (5):

$$AM = 30 + \frac{18{,}218}{41{,}215}\ 10 = 34{\cdot}42 \text{ years.}$$

The total of col. (6), divided by 41,215 and multiplied by 10^2 (to turn from units of 10 years to years, each squared), is the mean of squares of deviations measured from 30 years. The variance, or standard deviation squared, has deviations from $AM = 34{\cdot}42$ and not from the arbitrary point of 30 years. It can be shown that the mean of squared deviations from AM is the mean of squared deviations from an arbitrary point *less* the square of AM measured from the arbitrary point. Hence:

$$\begin{aligned} \text{Variance} &= \frac{185{,}637}{41{,}215}\ 100 - (34{\cdot}42 - 30)^2 = 450{\cdot}39 - 19{\cdot}54 \\ &= 430{\cdot}85 \\ SD &= \sqrt{430{\cdot}85} = 20{\cdot}76 \text{ years.} \end{aligned}$$

The variance calculated from grouped data in this way is easily seen to be in error, in general on the high side. If the frequency distribution is of the common type illustrated in Fig. 15, values of the variable in any class tend to cluster nearer the central peak of the distribution than the centre of the class. Hence, the deviations from the mean (which is near the peak) tend to be too large if all values are taken at the class centre. An approximate adjustment has been devised for cases where (*a*) the distribution is approximately symmetrical with a central peak, and (*b*) the classes are of equal length h. This is known as *Sheppard's correction:*

$$\text{Variance (corrected)} = \text{Variance (crude)} - \frac{1}{12}\ h^2$$

The correction is *not* applicable when the distribution is very skew (e.g., J-shaped). Nor is it applicable when the classes are of unequal length. It can be applied only roughly in the present case where the distribution is rather skew and all classes are of 10 years except the two end ones:

$$\text{Variance (corrected)} = 430{\cdot}85 - \frac{100}{12} = 422{\cdot}52$$

$$SD \text{ (corrected)} = \sqrt{422{\cdot}52} = 20{\cdot}56 \text{ years}$$

The calculation can be repeated for the population distribution of S.W. England, and medians and quartiles can be calculated on the assumption that the population is uniformly distributed in each class (5·3). The results are:

	AM (years)	SD (years)	M (years)	QD (years)	Skewness ^{1}
England and Wales	34·4	20·6	32·6	16·5	6·5
S.W. England	36·7	21·5	35·0	17·5	6·5

^{1} $\frac{1}{2}(Q_1 + Q_3) - M$ as percentage of QD.

It follows that the population of S.W. England is definitely older, on the average by more than two years, than in the country as a whole, but that there is little difference between the spread and skewness of the age distributions.

The method of calculating mean and standard deviation can be further illustrated with the frequency distribution of Appendix I, Table 13, as represented in Fig. 10 (3·7 above). The difficulty here is that the distribution is J-shaped with classes of very unequal length. The calculation is very approximate and Sheppard's correction to the variance is not applicable. For mines with output per manshift of under 15 cwts., the calculations are:

No. of wage-earners employed	Centre of class		No. of mines	Product	
	No.	origin: 374½ unit: 25		(3) × (4)	(3) × (5)
(1)	(2)	(3)	(4)	(5)	(6)
1–19	10	− 14·58	173	− 2,522·3	36,775·1
20–49	34½	− 13·6	39	− 530·4	7,213·4
50–99	74½	− 12	31	− 372	4,464
100–249	174½	− 8	33	− 264	2,112
250–499	374½	0	70	—	—
500–749	624½	10	42	420	4,200
750–999	874½	20	16	320	6,400
1,000–1,499	1,249½	35	16	560	19,600
1,500–1,999	1,749½	55	8	440	24,200
2,000–2,499	2,249½	75	1	75	5,625
Total			429	− 1,873·7	110,589·5

$$AM = 374{\cdot}5 - \frac{1{,}873{\cdot}7}{429}\ 25 = 265{\cdot}3$$

$$\text{Variance} = \frac{110{,}589{\cdot}5}{429}\ 625 - (265{\cdot}3 - 374{\cdot}5)^2 = 149{,}194$$

$$SD = \sqrt{149{,}194} = 386{\cdot}3$$

5.9 *Weighted Averages.* A somewhat different problem of frequent occurrence is illustrated by the data of Table 8 of Appendix I. The wheat yield, in cwts. per acre under wheat, is given in each of twenty-two counties in S.E. England in 1929–38. What is the wheat yield for the whole area? Clearly some kind of average of the twenty-two individual figures; but what kind of average?

If the acreage under wheat in 1929–38 is given for each county, the correct procedure is to multiply each wheat yield by the corresponding acreage under wheat (equals output of wheat in each county), sum the products and divide by the total acreage under wheat in the whole area. The result can be viewed as a "weighted" average of the twenty-two separate wheat yields, each yield being multiplied by its appropriate "weight" (acreage under wheat), and the sum divided by the total of the "weights." This concept is a development of the ordinary arithmetic mean. In general terms, a set of quantities x is given, to each of which is attached a weight w, and the *weighted arithmetic mean* is obtained as $\frac{\Sigma wx}{\Sigma w}$. A corresponding weighted geometric mean can be defined, each quantity of x being raised to the power of w, the product obtained and the Wth root extracted (where W is the sum of the w's).

The weights w indicate the relative importance of the various x's. In the particular case where the items are of *equal* importance, each $w = 1$ and the weighted mean reduces to the ordinary mean. The weighted mean, then, allows for items of *unequal* importance. Further, the exact w's need not be known; any set of weights proportional to them will give the correct weighted mean. It is relative importance of items that matters. In the example of wheat yields the relative importance of each county's yield is shown by the acreage under wheat, the appropriate weights for the yields. But any numbers proportional to the acreages will serve equally well as weights.

The ordinary (unweighted or simple) mean is not appropriate since the counties are not of equal importance as regards wheat production.

Now suppose that the correct weights, here acreages under wheat, are not available. We can still obtain an *estimate* of the wheat yield in the whole area if we can determine a set of values approximately proportional to the unknown acreages under wheat and if we use these as weights. We derive various estimates by using different approximations to the correct weights. In practice, we find that the result varies little as long as the weights used are fairly close approximations. This is illustrated by the following calculations, the items corresponding to counties in the order of Table 8:

Yield (cwts.)	First weights	Second weights	Product (1) × (2)	Product (1) × (3)
(1)	(2)	(3)	(4)	(5)
17·7	238	35	4,212·6	619·5
16·4	314	30	5,149·6	492·0
17·1	365	25	6,241·5	427·5
16·6	251	50	4,166·6	830·0
22·0	205	50	4,510·1	1,100·0
18·3	676	100	12,370·8	1,830·0
16·8	279	45	4,687·2	756·0
16·8	192	40	3,225·6	672·0
20·2	617	40	12,463·4	808·0
17·7	446	20	7,894·2	354·0
18·2	31	0	564·2	0
19·0	938	120	17,822·0	2,280·0
16·9	484	40	8,179·6	676·0
17·4	41	5	713·4	87·0
16·4	380	40	6,232·0	656·0
18·1	88	5	1,592·8	90·5
19·3	428	60	8,260·4	1,158·0
18·0	284	50	5,112·0	900·0
16·2	177	10	2,867·4	162·0
17·2	300	15	5,160·0	258·0
17·5	221	15	3,867·5	262·5
16·7	459	25	7,665·3	417·5
390·5	7,414	820	132,958·2	14,836·5

The simple mean of the yields is 17·75 cwts., the sum of col. (1) divided by twenty-two. This is known to be inappropriate as a measure of the wheat yield for the whole area. Cols. (2) and (3) are approximate weights for the yields of the various counties. The first gives the total acreage under crops and grass (in 000 acres) in 1941, which will be very roughly proportional to the acreage under wheat in 1929–38. The second is an attempt at an improvement; 11 per cent of col. (2) is taken (this being known as the proportionate area under wheat in the whole region) and the result is adjusted up or down, and rounded off, according to general knowledge of the importance of wheat in the agriculture of the counties. From cols. (4) and (5), estimates of the wheat yield for S.E. England in 1929–38 are given by the weighted averages:

$$\text{First weights:} \quad \text{Yield} = \frac{132{,}958{\cdot}2}{7{,}414} = 17{\cdot}9 \text{ cwts.}$$

$$\text{Second weights:} \quad \text{Yield} = \frac{14{,}836{\cdot}5}{820} = 18{\cdot}1 \text{ cwts.}$$

The two sets of weights, though differing considerably in their proportions, give averages which are not far apart; indeed they are close enough for the conclusions that the wheat yield in S.E. England is 18 cwts. per acre. Even the unweighted (and inappropriate) mean is not much less.

Other examples of weighted averages can be drawn from transport statistics. Suppose the average length of haul is needed for all merchandise freight carried by the railways in a month. The unweighted mean of the distances the various consignments are carried is not appropriate since some consignments are of a few lb. and others of many tons of freight. An appropriate average is the weighted mean of the distances, each being weighted with the amount of freight (in lbs. or tons) in the consignment. This average, in fact, is what is calculated when the number of ton-miles of freight is divided by the total tonnage carried.

There are other applications of weighted averages where several averages with various weights can be used for different purposes. Appendix I, Table 7, gives the wholesale price of eggs (per 120) in each month of 1938. What is the average

price for the whole year? One average is the unweighted mean of the twelve monthly quotations. Another is obtained by weighting each month's price with the domestic production of eggs in the month, and a third by weighting with the number of eggs consumed in the month. Available data are not sufficient to give these two sets of weights in 1938. An approximation to either set can be obtained from information on national markings and on consumption of eggs in various periods. This is shown below in the form of rough percentages of annual production or consumption in the various months:

1938	Jan.	Feb.	Mar.	Apr.	May	June	July	Aug.	Sept.	Oct.	Nov.	Dec.
Price (x)	16/2	14/6	10/4	10/4	11/4	12 2	15 4	16/9	18 5	19 11	21/9	17/8
Weight (w, %)	6	8	11½	13	11½	10	8	7	7	6	6	6

Here $\Sigma x = 184/8$, $\Sigma wx = 1{,}432/6$, $\Sigma w = 100$. Hence:

$$\text{Unweighted mean price} = \frac{184/8}{12} = 15/5$$

$$\text{Weighted mean price} = \frac{1{,}432/6}{100} = 14/4$$

There is a big difference between the two means, since the months of heavy production are also those of low prices. Each mean has its appropriate interpretation and use. The unweighted mean of 15/5 is the average price (per 120 eggs at wholesale) the producer would get in the year if he marketed equal numbers of eggs each month, or that the consumer would pay if he had the same number of eggs each month. The weighted mean of 14/4 is an approximation to the average price the producers actually got in the year with supplies varying from month to month as they do. It is also an approximation to the average price that consumers pay in the year with consumption varying according to the season.

A further extension of the use of a weighted average can be made to the case where the items to be averaged are only loosely related so that no "correct" weights exist. Such is the problem of indicating the "general level" of prices or of production. This leads us to consider the whole question of index numbers.

CHAPTER VI

INDEX NUMBERS

6.1 *The Concept of Index Numbers.* The prices of six selected grades of coal are quoted by the Board of Trade as follows:

	S. Wales ordinaries f.o.b.	Durham gas f.o.b.	Fifeshire steam f.o.b.	Lancs best house at pit	S. Yorks hard at pit	Notts best brights at pit
Price quotations (s. per ton)						
(1) Aver., 1930	18·25	15 39	12 00	28 64	14·56	20 78
(2) Aver., 1934	18·25	14 42	12 82	27·43	16·51	20 70
(3) July, 1938	23 00	19 25	15 97	27·17	21·13	20 00
Price relatives (Aver., 1930 = 100)						
(4) Aver., 1934	100 0	93·7	106 8	95·8	113·4	99 6
(5) July, 1938	126 0	125 1	133 1	94 8	145·1	96·2
Price relatives (Aver., 1934 = 100)						
(6) Aver., 1930	100 0	100 7	93 6	104 4	88·2	100 4
(7) July, 1938	126·0	133 5	124 6	99 1	128 0	96·6
Price relatives (July, 1938 = 100)						
(8) Aver., 1930	79·3	79 9	75·1	105·4	68 9	103·9
(9) Aver., 1934	79·3	74 9	80 3	101·0	78·1	103·5

If we are satisfied to take one grade of coal as representative of the whole group, we can follow without difficulty the variation of coal prices over time. We find it convenient to express the price in each period as a percentage of the price in a given period selected for purposes of comparison. We obtain for each period a *price relative* on the given *base period* as 100. For example, if we take Durham gas coal as representative and 1930 as the base period, we find the price relatives are 93·7 in 1934 and 125·1 in July, 1938; the price fell by 6·3 per cent. from 1930 to 1934, but in July, 1938, it was 25·1 per cent above 1930. All the various price relatives for each of the six grades of coal are shown above.

However, we can scarcely expect that the price of one grade of coal is representative of all the various coal prices.

But we may be content to take a small and careful selection of coal prices, such as the six shown above. In this case how do we measure the general level of coal prices and its variation over time? The prices themselves cannot be averaged since a ton of Durham gas coal is not the same thing as a ton of Notts best brights. When the problem is extended to include prices of (e.g.) steel and timber, there is certainly no way of adding tons of coal, tons of steel and loads of timber. One solution of the problem turns on the derivation of price relatives; these are numbers independent of units and they can be averaged. Though we cannot write an average coal price, we can obtain an average change in the level of coal prices. The change in each coal price from 1930 to July, 1938, is shown by the figure of row (5) above; an average of these figures gives the general level of coal prices in July, 1938, relative to 1930. Such an average is called an *index number* of coal prices. The example we are using is, in fact, part of the index number of wholesale prices calculated by the Board of Trade.

A whole series of problems now emerges and can be considered under three heads. First, the six coal prices are only a selection from all the prices currently quoted. How are these prices obtained, are they recorded accurately, and are they representative of other coal prices? Such questions concern the *choice of items* in the index number and the sources of data available on them.

Next, if we decide to obtain an index by averaging the price relatives of the six grades of coal, what kind of average should we use? We can choose the arithmetic or geometric mean, the median or some other average and our result will vary accordingly. From rows (4) and (5) above, we find:

Average of price relatives (aver. 1930 = 100)			
Date	AM	GM	Median
Aver., 1934	101·5	101·3	99·8
July, 1938	120·1	118·6	125·6

These are unweighted averages and assume, in effect, that the six coal prices are all of the same importance in the base year 1930. Is this an appropriate assumption? Since the different grades of coal enter into consumption and export to varying amounts, why not use these amounts to weight the average

used in the index? Moreover, there may be other, and perhaps better, methods of obtaining an index than averaging price relatives. These problems are all concerned with the *choice of formula* for combining the data into an index number.

Thirdly, the selection of 1930 as a base period is arbitrary. What difference is there when another base is selected, and how is an index number switched from one base to another? Such problems concern the *choice of base period* which is selected for writing as 100 in the index number.

The problems have been posed in terms of a particular price index number. Historically, index numbers were first developed for prices, to measure the "general level of prices," and hence (inversely) the "purchasing power of money." Index numbers, however, are not confined to prices. Their range is very great and they can indicate changes in *price* such as wage-rates, shipping freights and security prices as well as commodity prices, in *volume* such as output or business activity and in *value* such as retail sales or profits.

The classical definition of an index number is provided by Edgeworth who states that it shows by its variations the changes in a magnitude which is not susceptible either of accurate measurement in itself or of direct valuation in practice.[1] An index number is an indirect measure of something, a statistical concept. It is like an average or a measure of dispersion and, once again, we must expect to get several alternative measures, and not just one, to serve our purpose.

6.2 *Choice of Items.* An index number is sometimes viewed rather broadly as an indicator of the general level of prices in an extensive group, such as the group of all wage rates or of all commodity prices at wholesale. To obtain records of all the thousands of individual prices in the group is not a practical proposition, and consequently some kind of selection must be made. The problem can be regarded as one of sampling and it is confused by what is known or assumed about the relative importance of the different prices from which a selection is to be made, and about the relations between these prices. The prices of different grades of domestic coal must

[1]F. Y. Edgeworth, "The Plurality of Index Numbers" (*Econ. Jour.*, 1925).

be closely related to each other, and rather less closely to prices of export coal, of steel, and of many other items. These are complicated questions which we cannot consider here.

There is, however, considerable doubt whether it is appropriate to look at index numbers in this broad way. Some critics, such as Keynes, would reject the viewpoint altogether.[1] A more definite problem is to design an index number to measure changes in a more limited set of prices as they affect a particular group of individuals; for example, an index of retail food prices for working-class families or of prices received by farmers. Index numbers of volume or value can be designed similarly, e.g., for the volume of output of manufacturing industry or for the profits of transport undertakings. The choice of items is here directed as much at complete coverage as at obtaining a "representative" selection.

The problem is usually tackled best in two stages, as can be illustrated by an index of food prices for working-class families. First, the whole set of prices is arranged into a relatively small number of groups, each group being as homogeneous as possible. Secondly, individual prices are selected within each group for inclusion in the index. The groups may be meat, fish, fruit, vegetables, and so on. The individual items selected will then be particular cuts and grades of meat, particular vegetables, etc. The selection of prices within a group is partly a matter of taking important items for which prices are readily available. The more weighty consideration, however, is to get prices which are "representative" in the sense that their movements are typical of the whole group. On the other hand, the groups themselves give complete coverage of total working-class food expenditure, and some groups are more important than others in the aggregate. The groups, in fact, relate to a particular aggregate, here the total food expenditure of working-class families, and they should be "weighted" according to importance. The selection of groups is to be considered in relation to the choice of the weighting system.

[1] J. M. Keynes, *A Treatise on Money*, Vol. I (1930).

6.3 *Choice of Formula.* This is a double problem involving the choice of type of average or aggregate and the choice of the system of weights. To isolate the first question, let us assume that we have a set of price relatives, on a given base period, which we regard as of equal importance. A simple or unweighted average then serves as the index number, and the choice lies between the several types of average. The median can be ignored if only because it is seldom used, as can the mode and other less common forms. The effective choice, in fact, is between the arithmetic and geometric means, both commonly used and giving different figures for the index.

As we have seen (5.6 above), the geometric is less than the arithmetic mean of a given set of (unequal) price relatives. Moreover, the difference tends to increase as the time-span of the relatives is widened. If we can imagine a "true" index of prices, then either the geometric mean must be biassed downwards, or the arithmetic mean must be biassed upwards, or both, as an indication of the "true" value. There is one criterion, the so-called "time-reversal test," which suggests that it is the arithmetic mean which is biassed upwards.

Between 1930 and July, 1938, the price (shillings per ton) of Durham gas coal increased from 15·39 to 19·25. The price relative (as a ratio not a percentage) is 1·25 calculated forwards and 0·80 calculated backwards, these corresponding as two aspects of the same change in price. An increase of 25 per cent is wiped out by a subsequent decrease of 20 per cent. Now $\frac{1}{0{\cdot}80} = 1{\cdot}25$ so that the backward price relative is the reciprocal of the forward one, and conversely. From the technical statistical point of view, it is desirable that the index number should have the same reversible property as each of the separate price relatives, i.e., that the index calculated forwards and the index calculated backwards, on the same formula and between the same two dates, should be reciprocals of each other. Irving Fisher shows that the geometric mean, but not the arithmetic mean, satisfies this test.[1]

We can test this conclusion with the data of 6.1. Comparing 1930 and 1934, from rows (4) and (6), we have the index numbers (as percentages):

[1]Irving Fisher, *The Making of Index Numbers* (1922).

		AM	*GM*
Forward:	1934 on 1930 = 100	101·5	101·3
Backward:	1930 on 1934 = 100	98·9	98·7

The geometric means are reciprocals, but the reciprocal of the backward arithmetic mean is $\frac{1}{0{\cdot}989} = 1{\cdot}011$ which is not equal to the forward mean 1·015. A similar result is found for index numbers comparing 1930 and July, 1938:

		AM	*GM*
Forward:	July, 1938 on 1930 = 100	120·1	118·6
Backward:	1930 on July, 1938 = 100	85·4	84·3

The gap between the arithmetic and geometric means is wider and so is the difference between the reciprocal of the backward arithmetic mean $\frac{1}{0{\cdot}854} = 1{\cdot}171$ and the forward mean 1·201.

Our conclusion is that, if we use an unweighted average for our index number, our preference is for the geometric mean which can be calculated forwards and backwards with equivalent results. As compared with the geometric form, the arithmetic has an upward bias which tends to be larger for comparisons over longer periods.

In introducing the problem of weighting, we pass to the more specific index number which is related to a definite aggregate. The problem can be illustrated by the calculation of an index number of retail food prices for working-class families. The comparison of Appendix I, Table 4, is between September, 1939, and July, 1914, and the price relatives are given in col. (5). The unweighted means are $AM = 140{\cdot}5$ and $GM = 137{\cdot}7$. Neither is appropriate since the sixteen items are by no means of the same importance to *working-class families*. More is spent on meat and bread, for example, than on fish or eggs. An obvious set of weights for the price relatives is relative expenditures on the items by the average working-class family at the base date (July, 1914) as given in pence in col. (7). The weights actually used by the Ministry of Labour in this case are shown in col. (1); they are proportional to the expenditure of col. (7) but rounded. Further, since the index is related to aggregate consumption, a weighted arithmetic (rather than geometric) mean is appropriate. The

calculation of the index of retail food prices in September, 1939 (July, 1914 = 100) is carried out in Table 4:

$$\text{Index} = \frac{\text{sum of col. (6)}}{\text{sum of col. (1)}} = \frac{46{,}524}{334} = 139{\cdot}3$$

Such an index number is a *base weighted* arithmetic mean of price relatives, the weights being expenditures in the base period. It is of perfectly general application.

The same index number can be obtained in a different way. The quantities of the sixteen foodstuffs, purchased at the base date (July, 1914), given in col. (2) of Table 4, constitute a fixed budget which can be valued at the prices ruling at any date. As long as the working-class family continues to buy (or is able to buy) these fixed amounts, its standard of living is maintained. From the computations in cols. (7) and (8) of Table 4, we see that the fixed budget cost 225·60 pence in July, 1914, and 314·63 pence in September, 1939. With the fixed budget as a basis, the change in this cost gives an index number of retail food prices in September, 1939 (July, 1914=100):

$$\text{Index} = \frac{\text{sum of col. (8)}}{\text{sum of col. (7)}} \times 100 = \frac{314{\cdot}63}{225{\cdot}60} \times 100 = 139{\cdot}4$$

This is the ratio of two *aggregates*, namely, the costs of the fixed budget at the two dates compared. It indicates the price change by specifying how much more money the family must have to buy as much of every item at the second date as at the first.

The aggregate form of the index, however, is identical with the base weighted average of price relatives. Consider the contribution of any item to the latter. The item contributes to the numerator a product, the expenditure on the item in the base period times the price relative, and this equals what has to be spent on the amount of the item bought in the base period but at the prices of the second period. Hence, the contribution is the same as in the numerator of the aggregate form. Similarly, the item contributes to the denominator of the weighted average the expenditure on the item in the base period (its weight), again the same as to the denominator of the aggregate form. The particular calculations above differ slightly in the first decimal place, but only because of rounding and approximating. On either computation the index cannot

be written more accurately than 139; retail food prices increased by 39 per cent from July, 1914, to September, 1939, on the basis of the budget of the average working-class family at the earlier date.

6.4 *Price and Quantity Index Numbers.* The data of Table 4, Appendix I, while including prices at both dates compared, give the quantities purchased only at the base date. This limits the choice of formula since an index based on the purchases at the second date cannot be derived. Nor can an index number of volume of consumption be calculated. These limitations are removed in the data of Table 3, Appendix I, where prices and quantities of exports are given at each date compared. Taking exports in 1938 in relation to those in 1935, we are able to write four valuations from Table 3:

	Valuation of exports in	At average values of	Gives value (£000)
(1)	1935	1935	9,766
(2)	1935	1938	10,089
(3)	1938	1935	13,378
(4)	1938	1938	13,824

Notice that values (1) and (4) are totals as recorded (but rounded) while values (2) and (3) are "cross" valuations obtained by summing approximate figures, and so subject to greater error (see 4.8 above).

Using the aggregate method developed for the food price index number, we have an index number of export prices (average values) in 1938 (1935 = 100):

First index = 1935 exports at 1938 prices as percentage of 1935 exports at 1935 prices

$$= \frac{\text{Value (2)}}{\text{Value (1)}} \times 100 = \frac{10{,}089}{9{,}766} \times 100 = 103{\cdot}3$$

This index, which is certainly not accurate to more than the decimal place shown, can also be derived as a weighted average of price relatives.[1] But a second index number can now be calculated. Instead of using exports in 1935 for the comparison of prices, we can use exports in 1938 to build up an

[1]In Table 3, obtain price relatives from cols. (4) and (5), and weight with the 1935 values of exports in col (7). Notice that the "all other" item needs to be included by assigning some arbitrary price change, very much as the item "fish" is handled in Table 4.

aggregate index. Hence, we have an index number of export prices in 1938 (1935 = 100):

Second index = 1938 exports at 1938 prices as percentage of 1938 exports at 1935 prices

$$= \frac{\text{Value (4)}}{\text{Value (3)}} \times 100 = \frac{13{,}824}{13{,}378} \times 100 = 103{\cdot}3$$

This index can also be interpreted as a weighted average of price relatives. Its reciprocal, in fact, is the same index as the previous one with the years 1935 and 1938 switched, and so the weighted average of price relatives (1935 on 1938 as 100) with values of exports in 1938 as weights. The two index numbers are basically of the same form, one being a base weighted average carried forward from the base year to the second year, and the other the reciprocal of the same average carried backwards from the second to the earlier year. It happens that the values of the two index numbers are the same (103·3); this is accidental and it is not true of other comparisons, e.g., 1946 compared with 1938.

These are index numbers of prices. The same four valuations suffice to give two corresponding index numbers of volume of exports. For the volume of exports in 1938 (1935 = 100), we have:

First index = 1938 exports at 1935 prices as percentage of 1935 exports at 1935 prices

$$= \frac{\text{Value (3)}}{\text{Value (1)}} \times 100 = \frac{13{,}378}{9{,}766} \times 100 = 137{\cdot}0$$

Second index = 1938 exports at 1938 prices as percentage of 1935 exports at 1938 prices

$$= \frac{\text{Value (4)}}{\text{Value (2)}} \times 100 = \frac{13{,}824}{10{,}089} \times 100 = 137{\cdot}0$$

Again, by accident, the two values are equal.

Table 3 also provides the data for a comparison of 1938 and 1946. Carrying through the calculations, and assembling the results, we have:

	Index 1938 (1935 = 100)		Index 1946 (1938 = 100)	
	1935 weights	1938 weights	1938 weights	1946 weights
Price index	103·3	103·3	161·2	156·4
Volume index	137·0	137·0	79·6	77·2

We can now interpret the changes in exports of beverages and cocoa preparations. From 1935 to 1938, the value of exports

rose from £9·8 millions to £13·8 millions; price changes were slight (about 3 per cent up) and practically the whole of the change in value was an expansion in volume of exports. From 1938 to 1946, the value of exports rose further to £17·3 millions; there was a large price rise of about 60 per cent, while the volume of exports declined by more than 20 per cent. We have succeeded in analysing the change in value of exports into the constituent price and volume changes. This is, in fact, the purpose of the index numbers.

6.5 *Laspeyre and Paasche Forms.* Of the pair of index numbers we have constructed (either for price or for volume), that with base weighting is called after Laspeyre, and the other after Paasche. To make the forms quite precise, we can use a simple algebraic expression which employs the summation notation of 5·4 above. Suppose the index combines n items. In the base period 0, let the n prices be p_0, p_0', p_0'', . . . $p_0^{(n)}$ and the n quantities q_0, q_0', q_0'', . . . $q_0^{(n)}$. Similarly, the prices and quantities in the second (or current) period 1 are denoted with a suffix 1. The summations used all cover the n different items so that, for example, Σp_0q_0 stands for $p_0q_0 + p_0'q_0' + p_0''q_0'' + \ldots + p_0^{(n)}q_0^{(n)}$, i.e., the total value of all items in the base period. Then the index numbers in the period 1 relative to the base period 0, expressed in ratio rather than percentage form, are:

Laspeyre price index:

$$P_{01} = \frac{\Sigma p_1q_0}{\Sigma p_0q_0} = \frac{\Sigma p_0q_0\left(\frac{p_1}{p_0}\right)}{\Sigma p_0q_0}$$

Paasche price index:

$$P_{01}' = \frac{\Sigma p_1q_1}{\Sigma p_0q_1} = \text{Reciprocal of } \frac{\Sigma p_1q_1\left(\frac{p_0}{p_1}\right)}{\Sigma p_1q_1}$$

Laspeyre volume index:

$$Q_{01} = \frac{\Sigma p_0q_1}{\Sigma p_0q_0} = \frac{\Sigma p_0q_0\left(\frac{q_1}{q_0}\right)}{\Sigma p_0q_0}$$

Paasche volume index:

$$Q_{01}' = \frac{\Sigma p_1q_1}{\Sigma p_1q_0} = \text{Reciprocal of } \frac{\Sigma p_1q_1\left(\frac{q_0}{q_1}\right)}{\Sigma p_1q_1}$$

Each form is expressed alternatively as a ratio of aggregates and as a weighted arithmetic mean of relatives.

Since the Paasche form is the reciprocal of the Laspeyre form worked backwards, these index numbers are "time reversible" if the Laspeyre and Paasche forms have the same value. As we have seen (6.3 above), it is a convenience from the point of view of statistical technique if the index numbers have this property. The unweighted geometric average has the "reversible" property, and so is preferred generally to the unweighted arithmetic average which does not have the property. It would, therefore, be convenient if $P_{01} = P_{01}'$ and $Q_{01} = Q_{01}'$. This is quite often the case, at least approximately, in practice. But it is by no means always true; sometimes P_{01} exceeds P_{01}', and sometimes P_{01} is less than P_{01}'. Moreover, there is no reason why we shall expect our statistical convenience to be met in this way. Certainly we would not reject an index number calculation (like the comparison of 1938 and 1946 exports of Table 3) simply because P_{01} and P_{01}' differ in value.

It must be emphasized again that the index numbers are designed to measure the price or volume change as it applies to a particular group of individuals in particular circumstances. As between the two periods compared, the group of individuals may be different and circumstances may change. The price index P_{01} looks at the price change from the point of view of period 0, P_{01}' from the point of view of period 1. These may easily be different, and each has its uses and interpretation. This is so whether we are considering the cost of living from the viewpoint of the average working-class family (as in Table 4), or the changes in export prices from the angle of the exporter (as in Table 3), or some other price comparison.

The relation between P_{01} and P_{01}' can be examined in various ways. One of them employs the concept of statistical "correlation" (in the sense of the following chapter); another proceeds through an economic definition of a "true" index of price changes. These topics have been extensively investigated in recent years, but we must here pause at the threshold and refrain from passing into the new domain.[1]

[1]*See*, for example, H. Staehle, "A Development of the Economic Theory of Price Index Numbers" (*Rev. Econ. Stud.*, 1935); R. Frisch, "The Problem of Index Numbers" (*Econometrica*, 1936); A. A. Konus and H. Schultz, "The Problem of the True Index of the Cost of Living" (*Econometrica*, 1939); J. R. Hicks, "Valuation of the Social Income" (*Economica*, 1940); A. L. Bowley, "Earnings and Prices, 1904, 1914, 1937–8" (*Rev. Econ. Stud.*, 1941).

Price and volume index numbers of Laspeyre and Paasche forms are quite simply related. If $V_{01} = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_0}$ denote the change in value, then

$$P_{01} \times Q_{01}' = P_{01}' \times Q_{01} = V_{01}$$

Hence, in "explaining" the change in value, we can associate the Laspeyre price form with the Paasche volume form, or conversely. Notice, also that Q_{01} must exceed Q_{01}' if P_{01} exceeds P_{01}' and similarly for a deficit.

Various "crosses" between P_{01} and P_{01}' have been suggested as practical index numbers of prices. One of them, recommended by Bowley and others, is obtained by averaging the quantities in the aggregate form of the index number:

$$\frac{\Sigma (q_0 + q_1) p_1}{\Sigma (q_0 + q_1) p_0}$$

Another, termed the "ideal" formula by Irving Fisher, is the geometric mean of P_{01} and P_{01}'. These "crosses" have no great advantage over P_{01} and P_{01}' themselves. If prices and quantities are both given at all dates compared, then it is best to calculate P_{01} and P_{01}' separately before proceeding. If quantities are given only at the base date, as often happens in practice, then P_{01} can and should be calculated, but not P_{01}' or any "crossed" form.

6.6 *Choice of Base Period.* There remains the very practical question of the choice of base period. A desirable property is that the index numbers on one base should be in proportion to those on a second base so that one series can be switched to the other by dividing through by a fixed factor, namely, the value of the original index at the date used as the second base. This property, similar to that of "time reversibility" would make the choice of base period quite immaterial.

From the data on coal prices of 6.1 above, we have the index numbers:

UNWEIGHTED ARITHMETIC MEAN FORM

	Direct Calculation		Based on 1930 with 1934 = 100
	1930 = 100	1934 = 100	
Aver., 1930	100·0	98·9	$\frac{100{\cdot}0}{101{\cdot}5} \times 100 = 98{\cdot}5$
Aver., 1934	101·5	100·0	$\frac{101{\cdot}5}{101{\cdot}5} \times 100 = 100{\cdot}0$
July, 1938	120·1	118·0	$\frac{120{\cdot}1}{101{\cdot}5} \times 100 = 118{\cdot}3$

UNWEIGHTED GEOMETRIC MEAN FORM

	Direct Calculation		Based on 1930 with 1934 = 100
	1930 = 100	1934 = 100	
Aver., 1930	100·0	98·7	$\frac{100{\cdot}0}{101{\cdot}3} \times 100 = 98{\cdot}7$
Aver., 1934	101·3	100·0	$\frac{101{\cdot}3}{101{\cdot}3} \times 100 = 100{\cdot}0$
July, 1938	118·6	117·0	$\frac{118{\cdot}6}{101{\cdot}3} \times 100 = 117{\cdot}0$

With the geometric mean, therefore, the index numbers on 1934 = 100 are the same whether calculated from price relatives on 1934 as 100 or by the proportionate method of translation from the index numbers on 1930 = 100. On the other hand, the two processes give different results for the index numbers using the unweighted arithmetic mean. The unweighted geometric mean has the desired property, but the unweighted arithmetic mean has not.

Hence, if the base period is changed in an index number of unweighted arithmetic form, the new series of index numbers is not in proportion to the old series. Or, if the simple proportionate method of translation is adopted, we are in fact using a different formula (not an unweighted arithmetic mean). The reason for this is not far to seek. The simple arithmetic

index number assumes that all items are of the same importance at the base date. Between two periods, relative prices change and so does the relative importance of the items. Hence, the items cannot be of the same importance at both the base dates.

Exactly the same difficulty is met with index numbers of Laspeyre and Paasche form. With three periods 0, 1 and 2, the Laspeyre price index numbers on period 0 as base are $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$ and $\frac{\Sigma p_2 q_0}{\Sigma p_0 q_0}$ Switching to period 1 as base, the index number in period 2 becomes $\frac{\Sigma p_2 q_1}{\Sigma p_1 q_1}$ which is not the same as dividing the original index numbers, i.e., not the same as

$$\frac{\Sigma p_2 q_0}{\Sigma p_0 q_0} \text{ divided by } \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$$

It follows that the choice of base period in such index numbers is a matter of some importance and affects the relative values of the index numbers at various dates. This should not, however, cause too much alarm since the differences are generally not great.

The "chaining" of index numbers, often employed in practice, is a particular case of the same difficulty. The usual, or fixed base, index number compares each date directly with the given base date. The chain method uses a comparison of one date with the preceding one and so, by multiplication, back to the base date. The process can be illustrated with the Laspeyre price index number and four dates 0, 1, 2 and 3. The index number in period 3 on the base period 0 by the fixed base method is $\frac{\Sigma p_3 q_0}{\Sigma p_0 q_0}$ and by the chain method it is $\frac{\Sigma p_1 q_0}{\Sigma p_0 q_0} \times \frac{\Sigma p_2 q_1}{\Sigma p_1 q_1} \times \frac{\Sigma p_3 q_2}{\Sigma p_2 q_2}$ which is not the same thing. If the nature of the available data makes it necessary to adopt the chain method in practice, it must be remembered that the resulting index number is not identical with, though it may not differ greatly from, that obtained by the direct comparison.

6.7 *Standardization.* Index numbers of the Laspeyre and Paasche forms are intended to eliminate from a recorded change

in value, the effect of the change in prices and so to show up the change in volume. Or they eliminate the effect of changing volume and show up the price change. The results can be expressed in a somewhat different manner which is sometimes to be preferred. The aggregate form of such index numbers consists of one value divided by another. The device now to be adopted is simply to show the numerator and the denominator separately.

The valuations of Table 3, Appendix I, on exports of beverages and cocoa preparations can be displayed as follows:

Year	Recorded Value (£000)	Standardized Value[1] (£000)	Volume Index (1938 = 100)
	(1)	(2)	(3)
1935	9,766	10,089	73·0
1938	13,824	13,824	100·0
1946	17,209	11,000	79·6

[1]At average values of 1938.

The index numbers of col. (3) are obtained by dividing through the value of col. (2) by 13,825. They are the Laspeyre index numbers of volume based on 1938 and using 1938 prices. The point now is that the standardized values of col. (2) serve equally well as indicators of volume and they can be turned into index number form by simple division. They represent what the value of exports each year would have been if the prices of 1938 had been used throughout; price changes are eliminated and the change in volume of exports disclosed.

This simple device can be profitably employed in many fields. The annual White Papers on *National Income and Expenditure* show personal consumption each year standardized by valuation at 1938 prices, indicating changes in the volume of consumption. If the figures are divided by the value of the consumption in 1938, the usual (Laspeyre) index numbers of volume of consumption on 1938 = 100 are obtained.

Other examples illustrate the fact that index numbers can be used for comparisons, not over time, but between different places or for different groups. Appendix I, Table 8, shows the rent in shillings per acre, averaged over all holdings, in each

of a group of counties. The rent depends, to some extent, on the size of the holding and this affects the comparison from one county to another with varying average size of holding. A standardized rent can be obtained for each county by calculating what the average rent would be if the holdings were distributed by size, not as actually found in the county, but as in England and Wales as a whole. The effect of varying size of holding is thus eliminated. The standardized rents can be turned into index numbers by dividing by the rent for England and Wales; these would be county index numbers of rent with England and Wales rent set as 100. A similar standardization process can be applied to earnings per week in different industries or at different times in order to eliminate the effect of varying age and sex groupings of the workers on earnings.

6.8 *Standardized Mortality Rates.* The most familiar use of the device of standardization is in vital statistics, and particularly in the calculation of death rates. The method can be illustrated by a comparison of death rates in different areas of the country. One disturbing factor, which needs to be eliminated, is here the differing age and sex composition of the population from one area to another.

Appendix I, Table 11, shows that the death rate in S.W. England in 1938 was 12·9 per 1,000 of the population, as compared with a rate of 11·6 in England and Wales as a whole. Does it follow that S.W. England is a relatively unhealthy area? Surely not, since Cornwall, Devon and the other counties in the area contain renowned health resorts. Moreover, if the death rates are compared in each age group shown in Table 11, it is found that the rate is lower in S.W. England at every single age. It is clear then, that it is the older age distribution of the area, and not higher mortality, which makes the death rate larger in S.W. England. A standardized death rate is needed to eliminate the effect of the age distribution.

The method is to calculate the death rate which would have occurred in an area if the age distribution had been that of the country as a whole and not that actually found in the area. The computation proceeds by weighting the death rate in each age group in the area with the numbers in the population of

England and Wales in the age group. The figures of col. (8) of Table 11 are weighted with those of col. (1):

Age (years)	England and Wales population (000's)	S.W. England death rates (per 1,000)	Product (1) × (8)
	(1)	(8)	
0–	2,818	12·9	36,352
5–	6,025	1·2	7,230
15–	6,572	2·1	13,801
25–	6,817	2·4	16,361
35–	6,034	3·7	22,326
45–	5,134	8·0	41,072
55–	4,240	17·3	73,352
65–	2,561	40·6	103,977
75–	1,014	119·0	120,666
Total	41,215		435,137

The standardized death rate in S.W. England is then

$$\text{S.D.R.} = \frac{435{,}137}{41{,}215} = 10{\cdot}6 \text{ per } 1{,}000$$

Whereas the "crude" death rate (12·9) exceeds the death rate (11·6) in England and Wales, the standardized death rate (10·6) is lower. The standardized rate can be turned into an index number by dividing by the death rate in England and Wales, giving an index of mortality of ninety-one (England and Wales = 100). These figures for S.W. England are comparable with similar figures for other areas. In South Wales, for example, the standardized death rate in 1938 is 13·9 per 1,000, giving an index of mortality of 120 (England and Wales = 100).

6.9 *Some Index Numbers in Practice.* Some index numbers computed in practice measure changes in a money value or aggregate which cannot be calculated directly or continuously. Examples are index numbers of the value of retail sales, stocks, profits and pay rolls. Most index numbers in the economic field, however, are intended to measure changes in prices or in volume and such index numbers naturally come in pairs, an index of price and an index of volume together accounting for changing money value. Of the pair, the price index is

occasionally, and the volume index rather more frequently, not computed in practice. The range of types is very great as is seen from the following short account of the more important economic index numbers computed in the U.K.

Agricultural Prices and Output. The Ministry of Agriculture issues an index intended to show changes in the level of prices as received by farmers, of crops and livestock of fixed quality. It is weighted with aggregate farm output and the formula is the modification of the Paasche form. It is published in the monthly *Journal* of the Ministry and in the annual *Agricultural Statistics*. No index of agricultural output is available apart from that included (before 1939) as a component of the index of production of the London and Cambridge Economic Service.

Wholesale Prices. The official index of wholesale prices is issued monthly in the Board of Trade *Journal*. The form is an unweighted geometric mean of price relatives of 200 items. Another index is published monthly in the *Statist*, continuing Sauerbeck's calculations from 1846; it is an unweighted arithmetic mean of price relatives of forty-five items (*see* Appendix I, Table 9). In these index numbers, apparently unweighted, a system of weights is implicit in the selection of items. Other index numbers of wholesale prices, of varying coverage, are compiled, e.g., by the *Economist* and *The Times* newspapers.

Production. Index numbers of production were computed quarterly before 1939, both by the Board of Trade and by the London and Cambridge Economic Service. The latter index number was resumed, on a monthly basis, in the Service's *Bulletin* for February, 1948. The official index number is now computed monthly by the Central Statistical Office and is published in the *Monthly Digest of Statistics* (1948).

External Trade. Index numbers of the volume and price (average value) of external trade are issued quarterly in the Board of Trade *Journal*. They are of aggregative form, Laspeyre's index for volume and Paasche's for price (*see* Appendix I, Table 3, and 6.4 above). The Board of Trade also issues, monthly, in the *Journal*, index numbers of import and export prices based on a selection of items and using a modification of Paasche's form.

Retail Prices and Consumption. Monthly index numbers of retail prices of food and other items purchased by working-

class families are published in the Ministry of Labour *Gazette*. Until June, 1947, the index (known as the Cost of Living Index) was a base-weighted mean of price relatives of Laspeyre form with July, 1914 = 100 (*see* Appendix I, Table 4, and 6.3 above); an interim price index is now in use. There is no corresponding index of working-class consumption by volume. The White Paper on *National Income and Expenditure* includes an annual index of the volume of consumption on an over-all basis and the corresponding price index can be computed from the data shown. The method followed is similar to that adopted for external trade data. The index of retail sales issued monthly in the Board of Trade *Journal* relates to the value of sales and does not distinguish price and volume changes.

Wage Rates. Monthly index numbers of wage rates are compiled by the Ministry of Labour and the London and Cambridge Economic Service on similar lines. They are intended to show changes in the level of full-time weekly wages of workers of unchanged grade, and the Laspeyre form is used with the aggregate wage bill in different groups at the base date as weights.

Security Prices and Yields. The Actuaries' Investment index, computed monthly and summarized in the *Economist*, is a recognized index of security prices, using a wide range of securities as a guide to investors. It is an unweighted geometric mean of price relatives, with an implicit weighting again involved in the selection of quotations. The *Financial Times* publishes daily index numbers of prices of thirty ordinary and twenty fixed interest stocks and of yields of ordinary shares. There are other index numbers for varying purposes.

Vital Statistics. Death rates are standardized by eliminating the effect of changing age composition of the population both for an area-by-area comparison (age composition of the whole country as standard) and for a temporal comparison (age composition in 1901 as standard). To overcome the difficulty of base weighting with a remote year as base, a "cross" weighted index is now used for short-run comparisons (Registrar-General's *Statistical Review* for 1941). This *comparative mortality index* is of the form $\frac{\Sigma (q_0 + q_1)\, p_1}{\Sigma (q_0 + q_1)\, p_0}$ where the q's are numbers in age and sex groups of the population, and the p's

are corresponding death rates. Birth rates are also standardized to give measures of fertility. The process is more involved and leads to a *net reproduction rate*. This is essentially a weighting of birth rates among women of different ages, with a "survival" table used as weights to allow for mortality as well as fertility. Measures of morbidity or sickness are being developed on the same lines.

CHAPTER VII

CORRELATION

7.1 *Scatter Diagrams.* So far, our comparisons have involved only a single character, such as family income. The range of the analysis is greatly extended when a second character is introduced, e.g., food expenditure as well as family income. The question then, is whether the two characters are related and, if so, in what way. These are the problems of correlation.

The simplest approach is in graphical terms. The value of one variable x and the associated value of a second variable y are given for each item in a group of n items. With suitable scales, values of x can be measured along a horizontal axis, and values of y along a vertical axis drawn on a sheet of graph paper. Each item is then represented by a single point on the graph, the point vertically above the mark on the horizontal axis given by the value of x for the item *and* horizontally level with the mark on the vertical axis given by the associated value of y. The whole group of items is shown by a *scatter diagram* of n points on the graph, from which the nature of the relation between the two variables can be seen. Appendix I, Table 8, provides simple examples of scatter diagrams of twenty-two points, one for each of the counties in S.E. England. The scatter diagram of Fig. 16A shows the relation of rents in 1941 to wheat yields in 1929–38, and that of Fig. 16B relates wheat and barley yields in 1929–38.

If the variables are independent, any value of one variable will tend to be associated equally with large and with small values of the other. The points will be spread over the scatter diagram as though thrown there at random. If one variable is determined uniquely from the other, the points of the scatter diagram will lie on some line or curve which represents this (perfect) relation between the variables. Between these extremes, the diagram will show a greater or less scattering of points according as the relation between the variables is weak or strong. For example, there is a rather weak correlation between

rent and wheat yield in Fig. 16A, and a stronger relation between wheat and barley yields in Fig. 16B.

This statistical concept of correlation is quite neutral as regards causation. One of the variables may be "caused" by the other, but this can only be known from other than statistical considerations. The scatter diagram and measures of correlation derived from it will say nothing on the matter. For example, from Fig. 16A we may suspect that crop yields in the past are a factor influencing the level of farm rents. But, as far as the diagram goes, the explanation may be the other

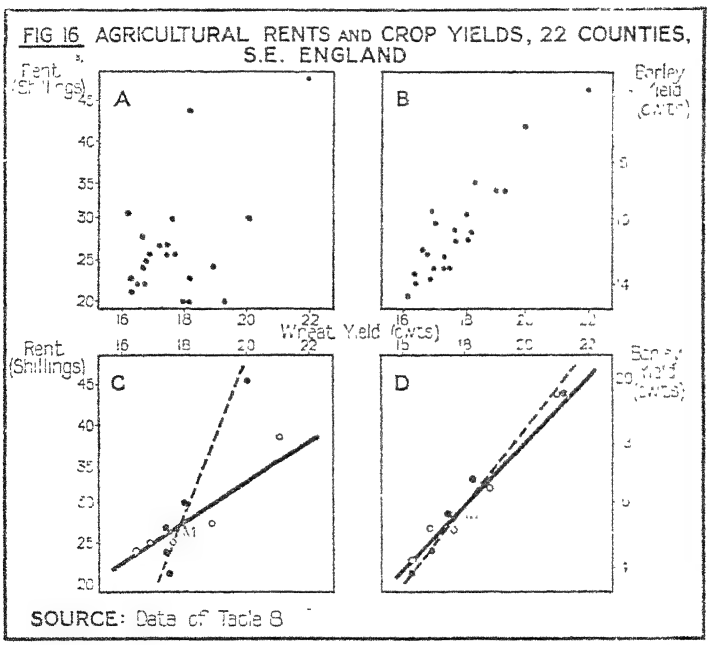


FIG 16 AGRICULTURAL RENTS AND CROP YIELDS, 22 COUNTIES, S.E. ENGLAND

SOURCE: Data of Table 8

way round; farmers paying high rents may be forced to cultivate intensively, and so get a high crop yield. Further, the relation may be indirect, both rent and crop yield being affected by other factors such as the size of holding.

7.2 *Regression Lines.* The first steps in the analysis of correlation can be taken on the scatter diagram. In Fig. 16A the

relation sought is between x = wheat yield (cwts. per acre) and y = rent (shillings per acre) in the twenty-two counties of S.E. England. There are twenty-two values of x from which the arithmetic mean is computed as 17·75 cwts. Similarly, the twenty-two values of y have mean 26·7 shillings. The two means are shown as a point M, the mean or central point of the scatter diagram. Next, convenient classes of x are taken, the corresponding values of y arranged in each class and the mean of them computed. So:

Wheat yield (cwts.)		No. of Counties	Rent (shillings)	Mean rent (shillings)
Range	Centre			
16·2–16·6	16·4	4	21, 22, 23, 31	24·25
16·7–17·1	16·9	5	22, 24, 25, 26, 28	25·0
17·2–18·1	17·65	7	20, 20, 26, 26, 27, 27, 30	25·1
18·2–19·6	18·9	4	20, 23, 24, 44	27·75
19·7–22·1	20·9	2	30, 48	39 0

The mean value of y (rent) is thus given for each of certain *arrays* of x (wheat yield) specified in the table. The means of arrays can be plotted on the scatter diagram against the central value of x in each array. Five points, shown in Fig. 16c as open circles, are here obtained for means of x arrays; they are to be read vertically as giving the mean rent in counties with wheat yield in the array concerned. The five points lie fairly close to a straight line passing through M and indicated in Fig. 16c by a solid line. This line, estimated graphically in Fig. 16c, is called the *regression line* of y = rent on x = wheat yield. The varying height of the line gives a simple but approximate relation of mean rent to any value of wheat yield which may be specified.

The process can be repeated for the relation of x = wheat yield to y = rent in the same data:

Rent (shillings)		No. of Counties	Wheat yield (cwts.)	Mean wheat yield (cwts.)
Range	Centre			
20–22	21	6	16·4, 16·6, 16·8, 18·0, 18·1, 19·3	17·5
23–25	24	5	16·4, 16·8, 16·9, 18·3, 19·0	17·5
26–28	27	6	16·7, 17·1, 17·2, 17·2, 17·5, 17·7	17·3
29–31	30	3	16·2, 17·7, 20·2	18·0
44–48	46	2	18·2, 22·0	20·1

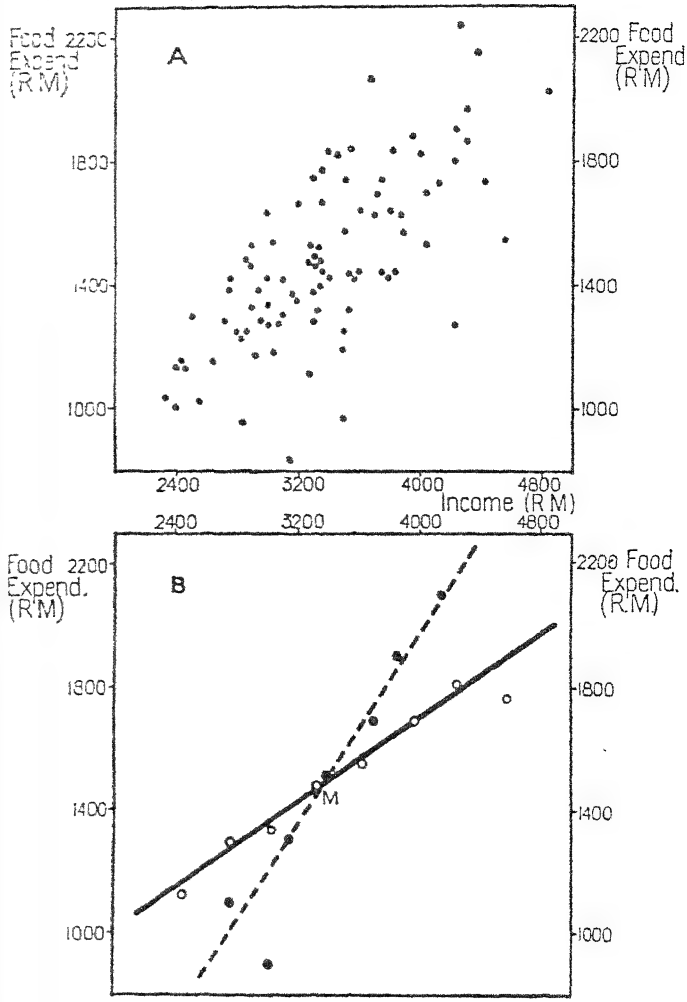
Five points, shown as solid circles, mark these array means in Fig. 16c, to be read horizontally as the mean wheat yield in counties with rents in the arrays concerned. The points are grouped about a straight line, an approximation to which is drawn in Fig. 16c. This is the regression line of x = wheat yield on y = rent.

Similar regression lines are drawn in Fig. 16D for the scatter of wheat and barley yields of Fig. 16B. An inspection of Fig. 16 now makes it clear that a measure of the degree of correlation between the variables is the angle between the regression lines. When the variables are strongly correlated, and the points of the scatter diagram cluster together, the regression lines lie close to each other (Fig. 16D). When the correlation is weaker and the points more scattered, the lines are farther apart (Fig. 16c).

The function of the regression lines, as approximate representations of means of arrays, is to isolate the mean value of one variable corresponding to any given value of the other; the variation of the first variable about its mean is ignored. A regression line is an *average* relation, and with it there is a *variation* of values about the average. In the regression of y on x, the variation ignored is in the vertical direction, a variation of y up and down about the line. In the regression of x on y, the variation ignored is a scattering of points to right and left about the line. The method here is similar to the analysis of a frequency distribution with measures of average and dispersion, but there are now two variables and so two directions of variation to be considered.

The present analysis is applicable to the case where means of arrays are located on a straight line, or approximately so—the case of *linear regression*. It is frequently found in practice. In other instances, less frequent but still commonly found, means of arrays lie approximately on curves rather than on straight lines and the regression is curvilinear. There is some indication of curvilinear regression in the relation of wheat yield to rent in Fig. 16c, and the representation by a regression line is rough. A more definite example of curvilinear regression is given in Appendix I, Table 13, where the mean size of coal mines first increases and then decreases as we pass from mines with low to mines with high output per manshift.

FIG. 17 FAMILY INCOME AND EXPENDITURE ON FOOD, HAMBURG & BREMEN 1927-28



SOURCE: Data of Table 10 A and D

Curvilinear regression must be analysed by methods different from those developed in the following sections.[1]

7.3 *An Example of Linear Regression.* Full details of income and food expenditure for each of a group of ninety families in Hamburg and Bremen in 1927–8 are given in Appendix I, Table 10A. The scatter diagram is drawn in Fig. 17. The correlation, as expected, is quite strong. Regression lines can be inserted by the method just described. Data of this type, however, are usually given in a double table as in Table 10D, and not in the full detail of Table 10A. The scatter diagram cannot then be drawn as in Fig. 17A. But such a table automatically specifies arrays of the two variables, the rows and columns of the table. From these, means of arrays and hence regression lines can be derived.

The over-all means, from the "borders" of the table as ordinary frequency distributions, are 3,386 RM for income and 1,484 RM for food expenditure. They give the mean point M through which the regression lines pass. From the columns of the table, each as a frequency distribution:

Centre of income array	2,450	2,750	3,050	3,350	3,650	3,950	4,250	4,550
Mean food expenditure	1,130	1,300	1,340	1,480	1,575	1,700	1,825	1,770

Similarly, from the rows of the table:

Centre of food expenditure array	900	1,100	1,300	1,500	1,700	1,900	2,100
Mean income	3,050	2,780	3,120	3,420	3,710	3,880	4,175

Each figure here is rounded to the nearest 10 RM. The means of arrays are plotted in Fig. 17B and approximate regression lines inserted. There is only a small number of families in this sample, but the variation of means of arrays about the regression lines is not large. Moreover, the lines are not far apart. It can be concluded that there is a close relation between food expenditure and income, and that the average relation of one variable to the other is approximately linear.

The exact meaning and use of regression lines must be kept in mind. The regression of food expenditure on income, the more interesting of the two here, isolates the relation food expenditure has to income on the average in this particular

[1] *See* Croxton and Cowden, Appendix II, Ref. (6), Chap. XXIII.

group of families. It ignores the variation of food expenditure from one family to another with the same income, the variation seen in the full scatter of Fig. 17A. The food expenditure of any one family is split into two parts. One part is given by the regression line, the part "explained" by the income of the family. The other part is peculiar to the particular family, the "unexplained" part which makes the family different from others with the same income. The first family of Table 10 has income 2,311 RM. According to the regression line of Fig. 17B, the food expenditure of a family with this income would be about 1,100 RM, on the basis of the average relation for the whole group of families. The actual food expenditure is 1,034 RM, i.e., this particular family spends some 66 RM less on food than the "norm."

Since the regression is taken as linear, the average relation of food expenditure to income is easy to interpret. On the average an increase in income is associated with a fixed proportionate increase in food expenditure. The proportion is given by the gradient of the regression line. As a rough estimate from Fig. 17B, food expenditure increases by about 35 RM, on the average, for every rise of 100 RM in income.

7.4 *The Correlation Coefficient.* So far, regression lines have been estimated roughly and graphically. It remains to make them more precise, to obtain actual measurements of correlation and regression. The problem can only be handled adequately in algebraic terms. The arithmetic mean and standard deviation are appropriate measures of average and dispersion; a corresponding measure of correlation is needed. The following is a convenient notation in the shorthand of algebra.

A group of n items is given, each with the associated values, x and y, of two variable characters. The n values of x by themselves make up a frequency distribution with

$$\bar{x} = \frac{1}{n}\Sigma x \qquad \text{Var } x = \frac{1}{n}\Sigma(x-\bar{x})^2 \qquad \sigma_x = \sqrt{\text{Var } x}$$

as mean, variance and standard deviation. Similarly, the frequency distribution of n values of y by themselves gives

$$\bar{y} = \frac{1}{n}\Sigma y \qquad \text{Var } y = \frac{1}{n}\Sigma(y-\bar{y})^2 \qquad \sigma_y = \sqrt{\text{Var } y}$$

By analogy with the variance, the mean of products of deviations of the two variables can be defined as the *covariance*:

$$\text{Cov } (xy) = \frac{1}{n} \Sigma (x - \bar{x}) (y - \bar{y})$$

The magnitude of the covariance indicates the extent of correlation between the variables. If there is little or no relation between the variables, a given value of one variable is associated equally with large and small values of the other; given a deviation $(x - \bar{x})$, the corresponding deviation $(y - \bar{y})$ is just as likely to be positive as negative. The products in Cov (xy) will tend to cancel out, some being positive and some negative. Cov (xy) is small if there is little correlation and, similarly, it is large if the correlation is strong. The sign of Cov (xy) is also important. There is *positive correlation* if large values of x and y are associated, and if small values of the variables likewise go together; Cov (xy) is then positive. There is *negative correlation* if large values of one variable go with small values of the other, i.e., positive deviations of one variable with negative deviations of the other, so that Cov (xy) is negative. Cov (xy) is measured in the units of the two variables themselves. It is often better to have a coefficient independent of units. One can be obtained by dividing Cov (xy) by the standard deviation of each variable, giving the *correlation coefficient*:

$$r_{xy} = \frac{\text{Cov } (xy)}{\sigma_x \ \sigma_y} = \frac{\frac{1}{n} \Sigma (x - \bar{x}) (y - \bar{y})}{\sqrt{\frac{1}{n} \Sigma (x - \bar{x})^2} \ \sqrt{\frac{1}{n} \Sigma (y - \bar{y})^2}}$$

If the variables are independent, then $r_{xy} = 0$. If there is an exact linear relation between the variables, then $r_{xy} = +1$ for variables which increase together and $r_{xy} = -1$ when one variable increases and the other decreases. These are special and extreme cases for which the results stated are easily proved algebraically. In practice, r_{xy} can be positive or negative and with value anywhere between zero and unity. The sign indicates whether the correlation is positive or negative, and the magnitude shows the degree of correlation.

7.5 *Calculation of the Correlation Coefficient.* The computation of r is an extension of that for the mean and standard

deviation and the whole process can be set out in one table. With the data on twenty-two counties in Appendix I, Table 8, let x = wheat yield (cwts. per acre), y = rent (shillings per acre) and z = barley yield (cwts. per acre). By addition and division by twenty-two:

$$\bar{x} = 17{\cdot}75 \qquad \bar{y} = 26{\cdot}7 \qquad \bar{z} = 15{\cdot}75$$

The rest of the work can be set out:

$x-\bar{x}$	$y-\bar{y}$	$z-\bar{z}$	$(x-\bar{x})^2$	$(y-\bar{y})^2$	$(z-\bar{z})^2$	$(x-\bar{x})$ × $(y-\bar{y})$	$(x-\bar{x})$ × $(z-\bar{z})$
− 0·05	− 0·7	− 0·05	0·0025	0·49	0·0025	0·035	0·0025
− 1·35	− 3·7	− 1·65	1·8225	13·69	2·7225	4·995	2·2275
− 0·65	− 0·7	− 1·15	0·4225	0·49	1·3225	0·455	0·7475
− 1·15	− 4·7	− 0·45	1·3225	22·09	0·2025	5·405	0·5175
+ 4·25	+ 21·3	+ 4·55	18·0625	453·69	20·7025	90·525	19·3375
+ 0·55	− 3·7	+ 1·45	0·3025	13·69	2·1025	− 2·035	0·7975
− 0·95	− 4·7	+ 0·55	0·9025	22·09	0·3025	4·465	− 0·5225
− 0·95	− 2·7	− 1·45	0·9025	7·29	2·1025	2·565	1·3775
+ 2·45	+ 3·3	+ 3·25	6·0025	10·89	10·5625	8·085	7·9625
− 0·05	+ 3·3	− 0·35	0·0025	10·89	0·1225	− 0·165	0·0175
+ 0·45	+ 17·3	− 0·15	0·2025	299·29	0·0225	7·785	− 0·0675
+ 1·25	− 2·7	+ 1·25	1·5625	7·29	1·5625	− 3·375	1·5625
− 0·85	− 1·7	+ 0·15	0·7225	2·89	0·0225	1·445	− 0·1275
− 0·35	− 0·7	− 1·15	0·1225	0·49	1·3225	0·245	0·4025
− 1·35	− 5·7	− 1·35	1·8225	32·49	1·8225	7·695	1·8225
+ 0·35	− 6·7	− 0·35	0·1225	44·89	0·1225	− 2·345	− 0·1225
+ 1·55	− 6·7	+ 1·25	2·4025	44·89	1·5625	− 10·385	1·9375
+ 0·25	− 6·7	+ 0·45	0·0625	44·89	0·2025	− 1·675	0·1125
− 1·55	+ 4·3	− 2·15	2·4025	18·49	4·6225	− 6·665	3·3325
− 0·55	+ 0·3	− 0·85	0·3025	0·09	0·7225	− 0·165	0·4675
− 0·25	+ 0·3	− 1·15	0·0625	0·09	1·3225	− 0·075	0·2875
− 1·05	+ 1·3	− 0·65	1·1025	1·69	0·4225	− 1·365	0·6825
		Total	40·635	1,052·78	53·875	105·45	42·755
		Mean	1·847	47·854	2·449	4·793	1·943

$$\therefore \text{Var } x = 1{\cdot}847 \quad \text{Var } y = 47{\cdot}854 \quad \text{Var } z = 2{\cdot}499$$
$$\text{Cov } (xy) = 4.793 \qquad \text{Cov } (xz) = 1.943$$

The correlation coefficient between wheat yield and rent is

$$r_{xy} = \frac{4{\cdot}793}{\sqrt{1{\cdot}847}\ \sqrt{47{\cdot}854}} = 0{\cdot}51$$

and that between wheat yield and barley yield is

$$r_{xz} = \frac{1{\cdot}943}{\sqrt{1{\cdot}847}\ \sqrt{2{\cdot}449}} = 0{\cdot}91$$

As the scatter diagrams show, the latter is a strong correlation and the former is much weaker.

The same method of calculation can be applied to the data on ninety families in Table 10A. Here, x = income and y =

food expenditure (in RM per year) and

$\bar{x} = 3{,}377 \quad \bar{y} = 1{,}481 \quad \text{Var } x = 299{,}639 \quad \text{Var } y = 78{,}568$

$$\text{Cov } (xy) = 108{,}489$$

So $$r_{xy} = \frac{108{,}489}{\sqrt{299{,}639}\ \sqrt{78{,}568}} = 0{\cdot}71$$

There is a fairly strong correlation.

If the same data are given only in grouped form, as in Appendix I, Table 10D, then an *estimate* of the correlation coefficient can be obtained on the assumption that the items in each class are concentrated at the centre of the class. Thus, Table 10D has six families with income 2,300–2,599 RM, and with food expenditure 1,000–1,199 RM; the assumption is that all these families have income 2,449·5 RM and food expenditure 1,099·5 RM. In a short-cut method of calculation, select origins and units:

x (income)	origin 3,349·5	unit 300 RM
y (food expend.)	origin 1,499·5	unit 200 RM

With the method of 5·8, the bottom row of Table 10D gives

$\bar{x} = 3{,}386 \quad \text{Var } x = 291{,}650$

and the right-hand column of the table gives

$\bar{y} = 1{,}484 \quad \text{Var } y = 80{,}200$

For the covariance, Table 10D is set out:

Food Expenditure	Income							
	−3	−2	−1	0	1	2	3	4
−3	–	1	1	1	–	–	–	–
−2	6	1	2	1	1	–	–	–
−1	1	5	9	4	2	–	1	–
0	–	3	4	8	6	3	–	1
1	–	–	2	3	5	3	2	1
2	–	–	–	2	1	3	3	–
3	–	–	–	–	1	–	2	1

The variables are measured from their selected origins and in their selected units.[1] The entry of one cell in the sum of products

[1]The last class of each variable is assumed to be of the same length as the others.

used to give the covariance is the product of the two variables times the number of families in the cell. Thus, for the second cell in the first column, the entry is $6 \times (-3) \times (-2) = 36$. The complete set of entries is:

0	6	3	0	0	0	0	0
36	4	4	0	−2	0	0	0
3	10	9	0	−2	0	−3	0
0	0	0	0	0	0	0	0
0	0	−2	0	5	6	6	4
0	0	0	0	2	12	18	0
0	0	0	0	3	0	18	12

The sum is 152 which is to be divided by ninety, and then, to turn into the original units of RM, multiplied by $300 \times 200 = 60{,}000$. The result is the mean product of deviations from the arbitrary origins selected. The covariance is the mean product of deviations from $\bar{x}$ and $\bar{y}$. As with the variance (5.8 above), it can be shown that the covariance is the mean product from arbitrary origins *less* the product of the means measured from the same origins. In this case:

$$\text{Cov}(xy) = \frac{152}{90} \times 60{,}000 - (3{,}386 - 3{,}349{\cdot}5)(1{,}484 - 1{,}490{\cdot}5)$$

$$= 101{,}333 + 566 = 101{,}900$$

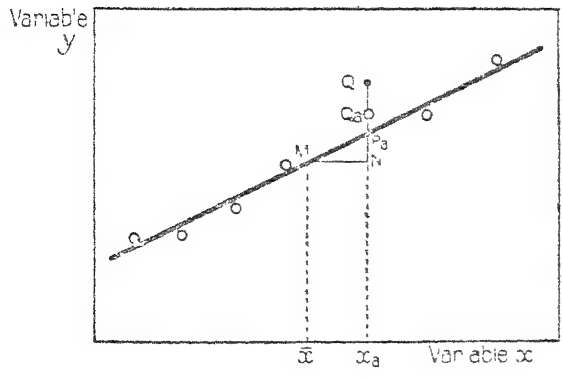
$$\text{So} \qquad r_{xy} = \frac{101{,}900}{\sqrt{291{,}650}\ \sqrt{80{,}200}} = 0{\cdot}67$$

This estimate can be compared with the correct value (0·71) obtained from the full data of Table 10A.

7.6 *Derivation of Regression Lines.* In 7.2 above, we inserted approximate regression lines in a scatter diagram, fitting them to means of arrays by a rough graphical process. We have still to define regression lines in precise terms. The regression of one variable y on another variable x is to be derived from the mean values of y in various arrays of x. With the notation of 7.4 above, let x_a define an array, usually as the central value of a range of x, and let n_a be the number of items and $\bar{y}_a$ the mean value of y in the array. The regression line is to give, for any x_a specified, a value of y as close as possible to y_a. But what precisely is meant by "as close as possible"?

Graphically, in Fig. 18, the problem is to draw a straight line through a set of points, marked with open circles, which represent means of arrays. Let Q_a be the mean of the array

FIG. 18 VARIATION FROM REGRESSION LINE



x_a and P_a the corresponding point on the straight line. For a line of "best fit," some of the Q_a's must lie above and some below the P_a's, and the deviations must balance. Positive (upward) deviations must cancel out negative (downward) deviations; the sum of all P_aQ_a's must be zero. This is, in fact, true of any straight line passing through the point M given by the over-all means of the variables. Hence, the "best fitting" line must be drawn through M, but its direction is still to be fixed. For this purpose, the size of the deviations P_aQ_a is to be considered without regard to whether Q_a lies above or below P_a. The deviations P_aQ_a, though cancelling out on balance, are smaller in magnitude for some lines than for others through M. To fix the line of "best fit" the deviations are to be as small as possible without regard to sign. The difficulty of the sign is avoided by squaring each deviation and then by making the sum of the squares as small as possible. The arrays have different numbers of items in them, and so

the deviations (squared) must be appropriately weighted. Hence, the condition for the line of "best fit" is that the sum of $n_a (P_aQ_a)^2$ over all arrays is a minimum. A regression so obtained is said to be fitted by "least squares" to array means.

Algebraically, the regression line is to give a value Y of y for any given value of x. It can be shown that the regression line of y on x as fitted by "least squares" is:

$$Y - \bar{y} = \beta_x (x - \bar{x}) \qquad \text{where } \beta_x = \frac{\text{Cov } (xy)}{\text{Var } x} = r_{xy} \frac{\sigma_y}{\sigma_y}$$

Similarly, the regression line of x on y is:

$$X - \bar{x} = \beta_y (y - \bar{y}) \qquad \text{where } \beta_y = \frac{\text{Cov } (xy)}{\text{Var } y} = r_{xy} \frac{\sigma_y}{\sigma_x}$$

Since $Y = \bar{y}$ when $x = \bar{x}$, and since $X = \bar{x}$ when $y = \bar{y}$, the two lines pass through the mean point M of the scatter diagram. Their direction is fixed by the constants β_x and β_y. The value of β_x, the *regression coefficient* of y on x, is the gradient of the regression line to the horizontal, i.e., the ratio of NP_a to MN in Fig. 18, a ratio constant for all points P_a on the line. The other regression coefficient is similarly interpreted in relation to the vertical axis.

The regression coefficients can also be interpreted as follows. The coefficient β_x can be positive or negative, which determines whether y increases or decreases, on the average, as x increases. Then, if the value of x is increased by 100 units, the value of y increases (or decreases) by 100 β_x units on the average. Another result of interest is that $\beta_x \times \beta_y = r_{xy}^2$; the correlation coefficient in value is the geometric mean of the two regression coefficients.

The computations of 7.5 give the regression coefficients and lines. For the regression of rent on wheat yield in the twenty-two counties of Appendix I, Table 8:

$$\beta_x = \frac{\text{Cov } (xy)}{\text{Var } x} = \frac{4{\cdot}793}{1{\cdot}847} = 2{\cdot}595$$

and the regression line is $Y - 26{\cdot}7 = 2{\cdot}6\ (x - 17{\cdot}75)$. So, on the average in these counties, an increase of 1 cwt. in wheat yield is associated with an increase of 2·6 shillings in rent per acre. Similarly, for the regression of food expenditure on income in the ninety families of Table 10A, the regression

equation is $Y - 1{,}481 = 0{\cdot}362\ (x - 3{,}377)$ and the same line from the grouped data of Appendix I, Table 10D, is $Y - 1{,}484 = 0{\cdot}349\ (x - 3{,}386)$. Hence, the average relation of food expenditure to income in this group of families is such that food expenditure rises by about 35 RM for every increase of 100 RM in income.

7.7 *Analysis of Variance.* As we have seen (7.3 above), the regression line serves to divide any given value of a variable into two parts, one being the part "explained" by a second variable by means of the regression and the other being an "unexplained" residual. To see the significance of this split, we need to examine how much of the total variation of the first variable about its mean is explained by the second variable and how much is left unexplained. The over-all variation of the variable is reduced to a smaller unexplained variation by finding and eliminating the part attributable to a known factor. The less is the amount left unexplained, the more significant and useful is the whole process. We can measure the variation of a character by its variance, and we have one application of what is called the analysis of variance. It is, however, easier to proceed with the sum of squares of deviations without dividing through by the number of items to get the variance.

The variable y is to be explained by a regression on the second variable x. With the notation of 7.6 and with Fig. 18 as illustration, the value of y for an item corresponds to the height of a point Q on the scatter diagram. The deviation of y from the over-all mean $\bar{y}$ is:

$$NQ = Q_aQ + P_aQ_a + NP_a$$

or
$$y - \bar{y} = (y - \bar{y}_a) + (\bar{y}_a - Y_a) + (Y_a - \bar{y})$$

Each deviation here must be taken with its proper sign. Now suppose squares of deviations are summed, starting with $\Sigma NQ^2 = \Sigma(y - \bar{y})^2$ over all items. It can be shown that this over-all sum of squares also breaks down into constituents:

$$\Sigma NQ^2 = \Sigma Q_aQ^2 + \Sigma P_aQ_a{}^2 + \Sigma NP_a{}^2$$

or
$$\Sigma(y - \bar{y})^2 = \Sigma(y - \bar{y}_a)^2 + \Sigma(\bar{y}_a - Y_a)^2 + \Sigma(Y_a - \bar{y})^2$$

The method of analysis can be illustrated with an example in detail. Appendix I, Table 14, shows a group of 484 families

distributed according to the number of rooms occupied (y) and to the number of persons (x) in the family. The variation of the number of rooms from one family to another is to be explained in part by a regression on the number of persons. Then, computations on the lines of 7.5 above give:[1]

$$\bar{x} = 3{\cdot}382 \quad \bar{y} = 2{\cdot}6095 \quad \text{Var } x = 3{\cdot}443 \quad \text{Var } y = 1{\cdot}395$$

$$\text{Cov } (xy) = 0{\cdot}9178 \quad \beta_{yx} = \frac{0{\cdot}9178}{3{\cdot}443} = 0{\cdot}2666$$

$$r_{xy} = \frac{0{\cdot}9178}{\sqrt{3{\cdot}443}\ \sqrt{1{\cdot}395}} = 0{\cdot}42$$

and the regression line is:

$$Y - 2{\cdot}6095 = 0{\cdot}2666\ (x - 3{\cdot}382)$$

The first array consists of seventy families with one person each:

No. of rooms (y)	No. of families	Product (1) x (2)	Deviation ($y - \bar{y}_a$)	Square of (4)	Product (2) x (5)
(1)	(2)	(3)	(4)	(5)	(6)
1	41	41	− 0·557	0·3102	12·72
2	21	42	0·443	0·1962	4·12
3	6	18	1·443	2·0822	12·49
4	2	8	2·443	5·9682	11·94
Total	70	109			41·27

The mean number of rooms is $\bar{y}_a = \frac{109}{70} = 1{\cdot}557$ from col. (3).

The sum of squares of deviations $\Sigma(y - \bar{y}_a)^2 = 41{\cdot}3$ from col. (6). The mean number of rooms given by the regression (for $x = 1$) is $Y_a = 2{\cdot}6095 + 0{\cdot}2666\ (1 - 3{\cdot}382) = 1{\cdot}974$. Hence,

$$\Sigma(\bar{y}_a - Y_a)^2 = 70 \times (1{\cdot}557 - 1{\cdot}974)^2 = 12{\cdot}2$$
$$\Sigma(Y_a - \bar{y})^2 = 70 \times (1{\cdot}974 - 2{\cdot}6095)^2 = 28{\cdot}3$$

Tabling these and similar results for other arrays:

[1]The class of eight persons and over is taken as nine persons, and that of six rooms and over as seven rooms.

Array (No. of persons)	No. of families	$\bar{y}_a$	Y_a	$(y-\bar{y}_a)^2$	$(\bar{y}_a-Y_a)^2$	$(Y_a-\bar{y})^2$
(1)	(2)	(3)	(4)	(5)	(6)	(7)
1	70	1·557	1·974	41·3	12·2	28·3
2	112	2·259	2·241	99·5	0·0	15·2
3	104	2·798	2·508	116·8	8·7	1·1
4	84	2·893	2·774	112·1	1·2	2·3
5	43	3·000	3·041	46·0	0·1	8·0
6	41	3·390	3·307	67·8	0·3	19·9
7	18	3·167	3·574	28·5	3·0	16·7
9	12	3·500	4·107	15·0	4·4	26·9
Total	484			527·0	29·9	118·4

The over-all variation in the number of rooms is

$$\Sigma(y-\bar{y})^2 = 484 \times \text{Var } y = 484 \times 1{\cdot}395 = 675$$

This is the sum of the cols. (5), (6) and (7) above:

$$\Sigma(y-\bar{y})^2 = \Sigma(y-\bar{y}_a)^2 + \Sigma(\bar{y}_a - Y_a)^2 + \Sigma(Y_a-\bar{y})^2$$

or $675 = 527 + 30 + 118$

This analysis of variance can be set out:

Variation	*Sum of squares*	
A Within arrays	$\Sigma(y-\bar{y}_a)^2$	527
B Array means from regression	$\Sigma(\bar{y}_a - Y_a)^2$	30
Total A + B		557
C Regression	$\Sigma(Y_a-\bar{y})^2$	118
Total A + B + C	$\Sigma(y-\bar{y})^2$	675

The sum of squares A is the variation of the number of rooms occupied apart from the effect of family size; it arises from individual differences between families of the same size and from many factors (e.g., income). The sum B is the variation of the array means from the regression, and it would be zero if the regression were exactly linear. The whole analysis depends on the regression being approximately linear, i.e., on B being small. The appropriate test is that B is small relative to A, which represents the "random" variation among families.

In the next stage, the sum (A + B) is taken as the total variation in number of rooms occupied apart from the regression, and compared with the sum C which is the variation due to family size (through the regression). In other words, of the total variation in number of rooms, C is the part explained by family size and (A + B) the part left unexplained. If the explanation of number of rooms occupied by means of family size is to be useful, C must be reasonably large relative to (A + B). In this instance, the explanation is of doubtful utility since C is rather small relative to (A + B). Precise tests are needed, however, and these are described in the technical literature on analysis of variance.[1]

An important result of the analysis can now be seen. Of the over-all variation in the number of rooms occupied, the proportion explained by family size is

$$\frac{C}{A + B + C} = \frac{118}{675} = 0{\cdot}175 = (0{\cdot}42)^2 = r_{xy}^2$$

The proportion left unexplained is $(1 - r_{xy}^2)$. This can be shown to be a perfectly general result.

The best interpretation of the correlation coefficient is in these terms. A variable y is to be explained by a second variable x. The total variation (Var y) divides up:

$$\begin{aligned} &\text{Variation explained by } x = r_{xy}^2 \text{ Var } y \\ &\text{Residual variation} \qquad\quad = (1 - r_{xy}^2) \text{ Var } y \end{aligned}$$

Hence, for most purposes, the *square* of the correlation coefficient is a better measure of the relation between x and y than the coefficient itself. In our example, the significant figure is $r_{xy}^2 = 0{\cdot}175$ rather than $r_{xy} = 0{\cdot}42$. Of the total variation in number of rooms occupied, $17\frac{1}{2}$ per cent is explained by family size and $82\frac{1}{2}$ per cent left as a residual to be explained by income and other factors.

7.8 *Relation between Laspeyre and Paasche Index Numbers.* It is possible to extend the concept of a weighted mean to weighted variance and covariance. One use of these is in the relation between index numbers of Laspeyre and Paasche

[1] *See* R. A. Fisher, Appendix II, Ref. (13), Chaps. VII and VIII, or Tippett, Appendix II, Ref. (17), Chaps. VI and VII.

forms. In the notation of 6.5 above, index numbers relating period 1 to period 0 are:

	Laspeyre form	*Paasche form*
Price index number	$P_{01} = \frac{\Sigma p_1 q_0}{\Sigma p_0 q_0}$	$P_{01}' = \frac{\Sigma p_1 q_1}{\Sigma p_0 q_1}$
Volume index number	$Q_{01} = \frac{\Sigma p_0 q_1}{\Sigma p_0 q_0}$	$Q_{01}' = \frac{\Sigma p_1 q_1}{\Sigma p_1 q_0}$

It can be shown[1] that the price index numbers are related:

$$P_{01}' - P_{01} = \frac{1}{\Sigma p_0 q_1} \Sigma p_0 q_0 \left(\frac{p_1}{p_0} - P_{01}\right) \left(\frac{q_1}{q_0} - Q_{01}\right)$$

and that a similar relation holds for $(Q_{01}' - Q_{01})$.

P_{01} is the weighted arithmetic mean of price relatives with the base values $(p_0 q_0)$ as weights. Q_{01} is similarly interpreted. Hence, the expression

$$\frac{1}{\Sigma p_0 q_0} \Sigma p_0 q_0 \left(\frac{p_1}{p_0} - P_{01}\right) \left(\frac{q_1}{q_0} - Q_{01}\right)$$

is the weighted covariance between price and quantity relatives over all items in the index numbers. From the relation above, $(P_{01}' - P_{01})$ is a positive multiple of this weighted covariance. So P_{01}' is larger than P_{01} if the (weighted) correlation between price and quantity relatives is positive, i.e., if price and quantity increases above average are associated; P_{01} is larger than P_{01}' if the correlation is negative, i.e., if a price increase above average is associated with a quantity increase below average. The correlation can be interpreted in terms of the economic relations of production and consumption in the two periods. It cannot be said that one index is always larger than the other (or that there is "bias") for the correlation found can change from one period of comparison to another.

[1]*See* A. L. Bowley, "Earnings and Prices, 1904, 1914, 1937–8" (*Rev. Econ. Stud.*, 1941).

CHAPTER VIII

TIME SERIES

8.1 *Analysis of Time Series.* We begin with a graph, drawn as in 4.2 to 4.3 above, from which we obtain our first impression of the variation in a time series and of the relation between one time series and another. Graphs can be misleading, however, and we need to subject our first impression to a closer scrutiny. We must develop more precise methods of analysis of time series. The variations of a time series are of many kinds which can be grouped under three heads. There is, first, the general direction of movement or the *trend* of the variable over the long period. Then there are *oscillations* of various types, of greater or less regularity, superimposed on the trend. Finally, there are *residual or irregular variations* which may arise from isolated events such as a war or general strike, or which may be due to the operation of random influences.

The isolation and interpretation of these three general components is the main task in the analysis of a single time series. The distinction between the components is not always clear-cut; it depends on the nature of the series and, in particular, on the over-all period covered and on the frequency of recording (e.g., monthly, quarterly or yearly). The trend is a long period movement relative to the length of the series. The trend in earnings in the period 1880–1914 (Table 5) is a longer period concept than the trend in egg prices over 1929–38 (Table 7). The trend over a period of ten years or less usually includes a movement which turns out to be an oscillation in the longer run of fifty years or so. There may be several different oscillations in a long time series, one on top of the other. As the length of the series is reduced, the longer oscillations do not have time to work themselves out and they become absorbed into the trend. The distinction between trend and oscillation is relative to the length of the series. Further, whether the shorter oscillations show up or not depends on the frequency of recording. An oscillation which works itself

out in a year or two will not be evident in a series given annually, but is seen when the recording is monthly.

The most common oscillation of the shorter kind is that which completes itself in a year and which is determined by seasonal factors. This is the *seasonal variation*, to be looked for only when the series is given quarterly or more frequently. Among longer oscillations, the economist is generally looking for one working itself out in a period of five to ten years, and perhaps also for others operating over longer periods. These are the oscillations associated with the well-known fluctuations or "cycles" of business activity.

The lumping together of irregular and random variations needs explanation. Such variations mean that any recorded figure in the series is different from what is expected from the trend and oscillations of the preceding period. It must not be thought that such variations occur and are then to be forgotten. On the contrary, they can cause a change in the existing trend and oscillations. For example, war in 1939 may be regarded as a residual factor for the trend and oscillations in a time series up to 1939, but a completely different trend and oscillation must be expected after 1945. This leads to the conclusion that the trend and oscillation in one part of a time series may be different from those in another part. In particular, the nature of seasonal variation can, and usually does, change from one period to another.

8.2 *The Method of Moving Averages.* The first step in the analysis of a time series is to isolate the trend appropriate to the over-all period considered. A simple method of general application is to use a moving average, and its success depends on the following considerations. Suppose that a series consists of a regular trend plus a single oscillation of perfect regularity operating in a period of n years. An average of the figures in any period of n consecutive years will not be influenced at all by the oscillation, high values exactly counterbalancing low values. The average, in fact, gives the trend value at the mid-point of the n years. The averages obtained for successive periods of n years will then describe the trend of the series. For example, if a series given annually has a regular oscillation every seven years, then the average of the years one to seven

inclusive will give the trend at year four (the mid-point of the first seven years); the average of years two to eight inclusive will give the trend at year five; and so on.

The presence of residual (random or irregular) variations in the series will usually make little difference; such variations tend to cancel out, some being up and some down, in any process of averaging. The main practical difficulty is that no oscillation is ever completely regular. The period of oscillation in the "cycle" of business activity can change and the spread (or amplitude) of an oscillation, including the seasonal variation, can vary in different periods. The method of moving averages, therefore, works only approximately in practice, and gives no more than an estimate of the trend. A period is first selected to represent approximately the length of the oscillation to be eliminated; a moving average of this length will then "smooth" the time series and give an estimate of the trend. There may be some doubt about the proper length to be selected, and it is usually desirable to experiment in practice with alternative moving averages, choosing that which appears to "smooth" the series most effectively. The average used may be the median, the arithmetic or the geometric mean. The median can be used effectively when the series contains occasional erratic values. The choice, however, lies usually between the arithmetic and geometric means with the former more commonly used.

The compilation of a moving arithmetic average is simplified by the technique shown in the following example. The trend is to be determined in the annual series of the cost of living from 1880 to 1914 (data of Table 5, Appendix I). Selecting a moving average of seven years, we derive cols. (2) and (3) from the original series in col. (1) of the table on the opposite page. Each entry in col. (2) is the sum of the seven items in col. (1) centred at the year in question. The first entry (689) is the sum of the first seven figures (1880–6) in col. (1). The second entry (672) is the sum of the seven figures in col. (1) from the second to the eighth inclusive (1881–7). When the first entry is written down, however, the second is obtained from it by the simpler process of adding on 88 (1887) and taking off 105 (1880):

$$\text{Second entry} = 689 + (88 - 105) = 689 - 17 = 672$$

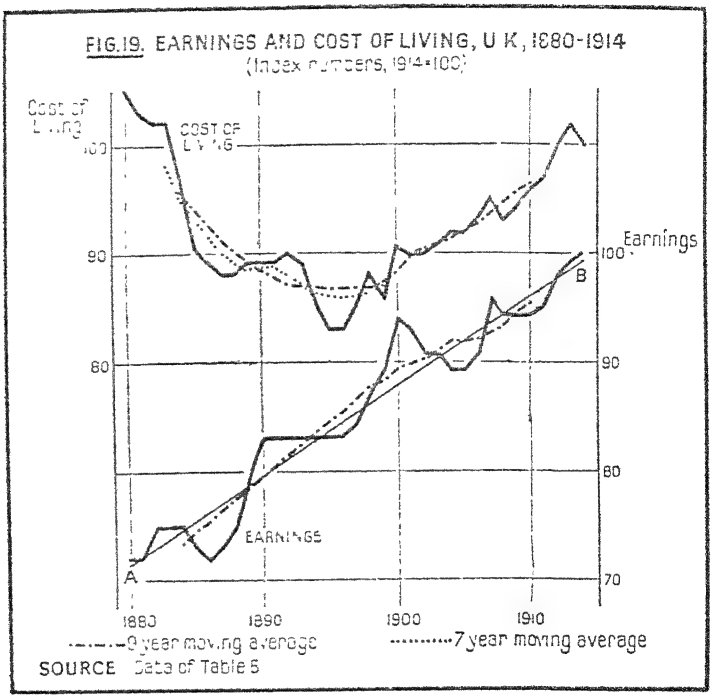
Year	Cost of living	Sums of 7's	Moving average (7 yrs.)	Sums of 9's	Moving average (9 yrs.)	Earnings	Sums of 9's	Moving average (9 yrs.)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1880	105					72		
1881	103					72		
1882	102					75		
1883	102	689	98·4			75		
1884	97	672	96·0	865	96·1	75	662	73·6
1885	91	657	93·9	849	94·3	73	670	74·4
1886	89	644	92·0	835	92·8	72	681	75·7
1887	88	631	90·1	822	91·3	73	689	76·6
1888	88	623	89·0	810	90 0	75	697	77·4
1889	89	622	88·9	802	89·1	80	705	78·3
1890	89	622	88·9	796	88·4	83	715	79·4
1891	89	619	88·4	790	87·8	83	726	80·7
1892	90	614	87·7	785	87·2	83	736	81·8
1893	89	608	86·9	782	86·9	83	745	82·8
1894	85	604	86·3	781	86·8	83	752	83·6
1895	83	603	86·1	778	86·4	83	758	84·2
1896	83	599	85·6	780	86·7	83	769	85·4
1897	85	601	85·9	780	86·7	84	779	86·6
1898	88	606	86·6	781	86·8	87	787	87·7
1899	86	613	87·6	787	87·4	89	795	88·3
1900	91	621	88·7	796	88·4	94	801	89·0
1901	90	628	89·7	805	89·4	93	807	89·7
1902	90	632	90·3	813	90·3	91	814	90·4
1903	91	639	91·3	820	91·1	91	823	91·4
1904	92	643	91·9	827	91·9	89	828	92·0
1905	92	646	92·3	830	92·2	89	828	92·0
1906	93	650	92·9	836	92·9	91	829	92·1
1907	95	655	93·6	843	93·7	96	833	92·6
1908	93	660	94·3	852	94·7	94	840	93·3
1909	94	668	95·4	862	95·8	94	850	94·4
1910	96	677	96·7	870	96·7	94	861	95·7
1911	97	682	97·4			95		
1912	100					98		
1913	102					99		
1914	100					100		

The third entry is then obtained from the second by adding on 88 (1888) and taking off 103 (1881):

Third entry $= 672 + (88 - 103) = 672 - 15 = 657$

The whole of col. (2) is thus derived entry after entry. To check the arithmetic, the last entry can be compared with the sum of the last seven figures in col. (1). The seven-yearly moving average in col. (3) is obtained by dividing each figure in col. (2) by seven.

A nine-yearly moving average can also be tried for this series, as computed in cols. (4) and (5). It remains to choose between the two moving averages, or to reject both in favour of another, as an estimate of the trend. Fig. 19 shows the original series of cost of living index numbers and the two moving averages on the same graph. Before 1900, it is seen



that the seven-yearly moving average does not get rid of the oscillations completely and that the other moving average gives a rather "smoother" trend. Further checking, e.g., with a moving average of eight or ten years, will confirm the nine-yearly average as the best estimate of the trend. A similar result is obtained for the series of earnings; Fig. 19 shows this series and its trend, computed in cols. (7) and (8) above as a nine-yearly moving average.

An alternative estimate of trend is given by a moving geometric average. This average attaches less weight to very large values of the variable (see 5.6 above) and the trend obtained by its use takes less account of large values which may occur in the time series. It is an appropriate trend when relative rather than absolute changes in the variable are important. The calculation of a moving geometric average is simplified by the fact that the logarithm of the geometric mean is the arithmetic mean of the logarithms of the original items. The moving geometric average is computed by first finding the moving arithmetic average of the logarithms of the items in the time series and then by looking up anti-logarithms.

8.3 *Elimination of Trend.* The trend, when computed, can be eliminated in one of two ways, by obtaining *either* the deviation of each original value from the value of the trend at the same date, *or* the ratio of the original to the trend value. The first is appropriate when actual variations (so much up or down), the second when relative variations (percentages above or below trend), are required. This is the same difference as between the use of natural and ratio scales.

The following is the procedure when a moving average defines the trend. If actual variations are important, the trend is computed as a moving arithmetic average, deviations from the trend give the series with trend eliminated and a graph on a natural scale is used. If relative variations are considered, a moving geometric average gives the trend, which is then eliminated by taking ratios to trend and a ratio scale is used in the graph. In practice, however, this logical separation is not always made, and a moving arithmetic average is sometimes used in both cases, as in the following example.

The table of 8.2 above gives the trends (nine-yearly moving averages) for the cost of living and earnings series from 1884 to 1910. The trends are eliminated by relating cols. (1) and (5) and cols. (6) and (8). For the cost of living in 1884, the deviation from trend is $97 - 96{\cdot}1 = +0{\cdot}9$, and the ratio to trend is $\frac{97}{96{\cdot}1} \times 100 = 100{\cdot}9$ per cent. The table is formed:

Cost of living			Earnings		Cost of living			Earnings	
Year	Dev. from trend	Ratio to trend	Dev. from trend	Ratio to trend	Year	Dev. from trend	Ratio to trend	Dev. from trend	Ratio to trend
1884	+ 0·9	100·9	− 1·4	101·9	1898	+ 1·2	101·4	− 0·7	99·2
1885	− 3·3	96·5	− 1·4	98·1	1899	− 1·4	98·4	+ 0·7	100·8
1886	− 3·3	95·9	− 3·7	95·1	1900	+ 2·6	102·9	+ 5·0	105·6
1887	− 3·3	96·4	− 3·6	95·3	1901	+ 0·6	100·7	+ 3·3	103·7
1888	− 2·0	97·8	− 2·4	96·9	1902	− 0·3	99·7	+ 0·6	100·7
1889	− 0·1	99·9	+ 1·7	102·2	1903	− 0·1	99·9	− 0·4	99·6
1890	+ 0·6	100·7	+ 3·6	104·5	1904	+ 0·1	100·1	− 3·0	96·7
1891	+ 1·2	101·4	+ 2·3	102·9	1905	− 0·2	99·8	− 3·0	96·7
1892	− 2·5	103·2	− 1·2	101·5	1906	+ 0·1	100·1	− 1·1	98·8
1893	+ 2·1	102·4	+ 0·2	100·2	1907	+ 1·3	101·4	+ 3·4	103·7
1894	− 1·8	97·9	− 0·6	99·3	1908	− 1·7	98·2	+ 0·7	100·8
1895	− 3·4	96·1	− 1·2	98·6	1909	− 1·8	98·1	− 0·4	99·6
1896	− 3·7	95·7	− 2·4	97·2	1910	− 0·7	99·3	− 1·7	98·2
1897	− 1·7	98·0	− 2·6	97·0					

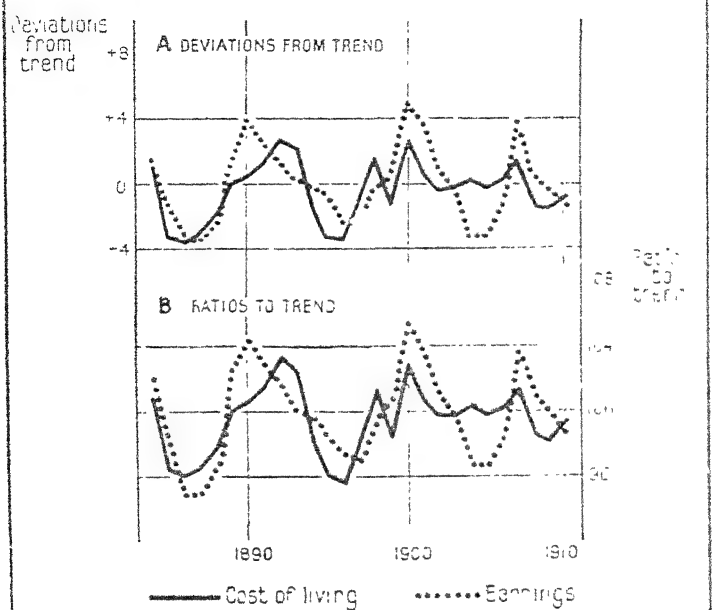
These series with trends eliminated are represented graphically in Fig. 20. Both series show oscillations, but those in earnings are more regular, varying from about 5 per cent below to about 5 per cent above the trend. The isolation of regular oscillations or "cycles" in trend-free series such as these has been the subject of much investigation. The problem is complicated, not so much because there are irregular variations left in the series, but rather by the fact that there may be several regular oscillations of different length superimposed one on the other.[1]

8.4 *Seasonal Variation.* One type of oscillation can be treated with more ease, the seasonal variation which shows up when the time series is given more frequently than annually. The simplification here is that the seasonal variation, if it exists at all, must have the definite period of one year.

Some general points must be stressed. First, an estimate of the seasonal movement can only be got after the trend in the original series has been removed. Otherwise, a February figure can be higher than January, not because of season, but because of an upward trend in the whole series. Secondly,

[1]The classical method of periodogram analysis has fallen into disrepute lately, particularly for economic data. A different method of approach, through serial correlation and autoregression, has been developed by Yule, Wold and Kendall (*see* M. G. Kendall, *Contributions to the Study of Oscillatory Time Series*, National Institute of Economic and Social Research, Occasional Papers IX, 1946).

FIG.20 EARNINGS AND COST OF LIVING U.K. 1884 1910



SOURCE: Data of Table 5

the amount of seasonal variation can change over time so that any estimate of it must be based on a definite and specified period. As we shall see, the estimate may be used outside the period on which it is based, but this involves the assumption that it continues substantially unchanged. Thirdly, any estimate of seasonal variation must be derived by an averaging process which gets rid of residual (irregular) variations. It cannot be found from one year alone, but only on the average over a number of years. Finally, the seasonal movement may be taken either as a deviation from trend or as a ratio to trend. In the former case, the seasonal movement is always the same amount in any one month, irrespective of the value of the time series in that month. In the latter case, the movement is of the

same relative amount, i.e., it grows as the trend of the series rises. A choice needs to be made and the appropriate procedures are as described in 8.3 above.

There is a great variety of methods of estimating seasonal variation.[1] They are, however, mainly variants of a basic process which is easily described, and which has three stages. The time series is assumed to be given monthly; the methods apply equally to other cases. The trend is first derived by moving averages or any other method. Then deviations of the original series from the trend are obtained and put in columns of months and rows of years. Finally, the January deviations are averaged, then the February deviations, and so on. The trend is eliminated at the first stage and the residual variations in the averaging process at the third stage. The final result is an average deviation for each month, representing the seasonal variation for the period considered.

The simplest form of the method is shown in the following computation for egg prices in the period 1929–38 (data of Table 7, Appendix I). The estimate of the trend, as a twelve-monthly moving average, starts as shown below:

Month	Price (pence)	Sums of 12 s	Average of pairs	Moving average (12 m.)	Month	Price (pence)
	(1)	(2)	(3)	(4)		(1)
1929 Jan.	252				1930 Jan.	226
Feb.	240				Feb.	216
Mar.	204				Mar.	140
April	140				April	134
May	155				May	131
June	155					
July	194	2,762	2,749	229	...	...
Aug.	221	2,736	2,724	227		
Sept.	235	2,712	2,680	223		
Oct.	315	2,648	2,645	220		
Nov.	341	2,642	2,630	219		
Dec.	310	2,618				

The sums of col. (2) are centred at mid-points of successive periods of twelve months, i.e., *between* months of the original series. Deviations from trend cannot be obtained until col. (3)

[1] *See* A. L. Bowley and K. C. Smith, "Seasonal Variations in Finance, Prices and Industry" (London and Cambridge Economic Service, *Special Memorandum*, No. 7, 1924).

is formed by averaging adjacent pairs of col. (2) and then placing the new entries against months. The trend is given in col. (4) on division by twelve. If the calculation is continued to the end, the trend of egg prices (in pence) is:

	Jan.	Feb	Mar.	Apr.	May	June	July	Aug	Sept.	Oct	Nov.	Dec.
1929							229	227	223	220	219	218
1930	216	214	212	209	205	200	196	193	191	190	188	186
1931	184	181	179	176	173	170	167	165	164	162½	162	161
1932	161	161	161	160	159	158	157	157	158	157	156	156
1933	155½	155	155	154	153	153	154	153	151	151	151	150
1934	150	150	149	148½	149	149	148	148	149	149	149	150
1935	151	152½	154	155	156	157	159	162	164	165	165½	166
1936	167	167	168	169	171	171	169	168	168	169	169	170
1937	172	173	175	176	177	179	182	184	184	185	186	187
1938	188	188½	189	190	189	186						

This represents the long-run trend plus the cyclical decline and recovery following 1929.

Next, deviations of the original prices from the trend are obtained by subtraction and written in rows and columns as shown below. These represent the normal seasonal movement plus all residual (irregular) variations in prices.

	Jan.	Feb.	Mar.	Apr.	May	June
	Price deviations from trend (pence)					
1930	10	2	— 72	— 75	— 74	— 59
1931	15	— 12	— 48	— 64	— 62	— 56
1932	13	— 14	— 49	— 56	— 59	— 44
1933	— 3½	12	— 46	— 63	— 57	— 39
1934	7	— 13	— 53	— 53½	— 57	— 38
1935	4	— 13½	— 53	— 63	— 54	— 38
1936	18	3	— 53	— 67	— 57	— 45
1937	— 17	— 9	— 45	— 68	— 61	— 38
(1) 8 *years*, 1930–7						
Totals	46½	— 44½	— 419	— 509½	— 357	— 481
Means	5·8	— 5·6	— 52·4	— 63·7	— 60·1	— 44·6
Seasonal variation[1]	6	— 5	— 52	— 63	— 60	— 44
(2) 4 *years*, 1930–3						
Totals	34½	— 12	— 215	— 258	— 252	— 198
Means	8 6	— 3·0	— 53·8	— 64·5	— 63·0	— 49·5
Seasonal variation[1]	10	— 2	— 52	— 63	— 62	— 48

	Jan.	Feb.	Mar.	Apr.	May	June
(3) 4 *years*, 1934–7						
Totals	— 12	— 32½	— 204	— 251½	— 229	— 159
Means	3·0	— 8·1	— 51·0	— 62·9	— 57·2	— 39·8
Seasonal variation[1]	3	— 8	— 51	— 63	— 57	— 40

	July	Aug.	Sept.	Oct.	Nov.	Dec.
	Price deviations from trend (pence)					
1930	— 19	1	13	82	107	55
1931	— 21	— 1	17	62½	109	36
1932	— 15	2	28	57	92	38
1933	— 24	7	25	48	91	59
1934	— 27	18	7	52	96	55
1935	— 18	20	16	42	79½	66
1936	— 17	12	18	76	77	56
1937	— 4	7	28	53	93	68
(1) 8 *years*, 1930–7						
Totals	— 145	66	152	472½	744½	433
Means	— 18·1	8·3	19·0	59·1	93·1	54·1
Seasonal variation[1]	— 18	9	19	60	93	55
(2) 4 *years*, 1930–3						
Totals	— 79	9	83	249½	399	188
Means	— 19·8	2·2	20·7	62·4	99·7	47·0
Seasonal variation[1]	— 19	3	22	63	101	48
(3) 4 *years*, 1934–7						
Totals	— 66	57	69	223	345½	245
Means	— 16·5	14·2	17·3	55·7	86·4	61·3
Seasonal variation[1]	— 17	14	17	55	86	61

1 Adjusted means.

The last stage is the averaging process which eliminates the residual variations. We fix the period of eight years (1930–1937) as the basis of the estimate of seasonal variation, shown in section (1) of the table above. The arithmetic mean of the eight January figures is found (5·8), then the arithmetic mean

of the eight February figures (— 5·6), and so on.[1] An adjustment is needed before these means can be taken as the seasonal variation. We require an estimate of the seasonal movement consisting of twelve deviations, some plus and some minus, balancing out to zero. In fact, the sum of the means is found to be — 5·1. Each mean is increased by $\frac{5 \cdot 1}{12} = 0 \cdot 4$ to the nearest first decimal place, and then rounded off to the nearest whole number. The line of adjusted means, which do add to zero, is our estimate of seasonal variation based on the period of eight years, 1930–7.

In illustrating the fact that estimates of seasonal variation vary as the period used is changed, we can compute the seasonal movement in egg prices in the four years 1930–3, and again in the four years 1934–7. The calculations are set out in sections (2) and (3) of the above table, and the two seasonal movements are shown in Fig. 21. The amplitude of the seasonal variation is not as great in the second period, when prices were generally lower, which suggests that the variation should be estimated in relative rather than in absolute terms.

The appropriate method of estimating relative seasonal variation is to vary the basic process by taking ratios to trend rather than deviations from trend and by using geometric means rather than arithmetic. The trend is derived as a moving geometric mean, ratios to the trend are written and averaged, month by month, with a geometric mean. The practical computations are not difficult. Logarithms of the original items in the time series are written, the basic (arithmetic) method applied to the logarithms, and the adjusted means which result converted back from logarithms to give the seasonal variation.

A mixed process is, however, sometimes employed in

[1]The arithmetic mean is the average appropriate to deviations from trend and it uses all the entries. Sometimes, however, we may find that a few deviations are so exceptional that we wish to give them less weight. If more than a few years are averaged, we could take the median of each month's deviations instead of the mean, or we could take a weighted mean in which the extreme deviations are given low weights (or even zero weights). Alternative averages of these kinds are quite often employed in practice.

practice, the basic arithmetic method being modified only by replacing deviations from trend by ratios to trend.[1] The mixed method is illustrated here in application to monthly egg prices in the four years 1934–7. The trend is given as above and the ratio of the original price to the trend is written in each month.

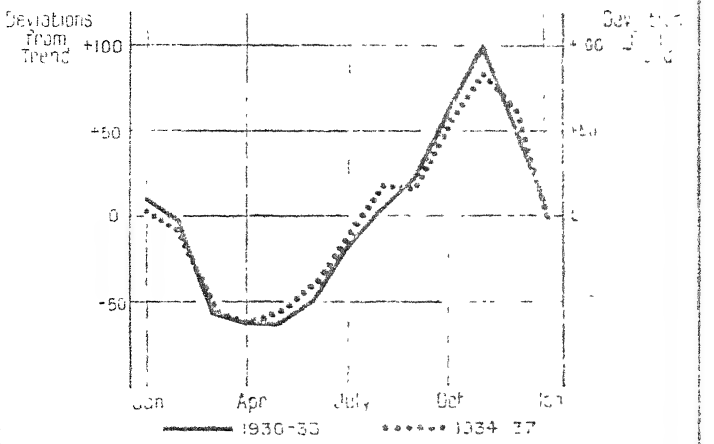
The ratio in January, 1934, is $\frac{157}{150} \times 100 = 104{\cdot}7$ per cent, i.e., the actual price is 4·7 per cent above the trend in the month. The whole series of ratios to trend is:

	Jan.	Feb.	Mar.	Apr.	May	June
			Price ratios to trend (%)			
1934	104·7	91·3	64·4	64·0	61·7	74·5
1935	102·6	91·1	65·6	59·4	65·4	75·8
1936	110·8	101·8	68·5	60·4	66·7	73·7
1937	90·1	94·8	74·3	61·4	65·5	78·8
Totals	408·2	379·0	272·8	245·2	259·3	302·8
Means	102·05	94·75	68·2	61·3	64·8	75·7
Seasonal variation	102·3	95·0	68·4	61·4	64·9	75·9
	July	Aug	Sept	Oct.	Nov.	Dec.
			Price ratios to trend (%)			
1934	81·8	112·2	104·7	134·9	164·4	136·7
1935	88·7	112·3	109·8	125·5	148·0	139·8
1936	89·9	107·1	110·7	145·0	145·6	132·9
1937	97·8	103·8	115·2	128·6	150·0	136·4
Totals	358·2	435·4	440·4	534·0	608·0	545·8
Means	89·55	108·85	110·1	133·5	152·0	136·45
Seasonal variation	89·8	109·1	110·3	133·8	152·3	136·8

The arithmetic mean of four trend ratios is obtained each month and adjusted to give an estimate of seasonal variation. A set of twelve figures is required varying around 100 as mean, and

[1] *See* 8.3. above. The logical geometric method is used by the London and Cambridge Economic Service (*Special Memoranda*, Nos. 7 and 36). The mixed method is employed, for example, in the index number of agricultural prices (*see* C. T. Houghton, "A New Index Number of Agricultural Prices," *Jour. Roy. Stat. Soc.*, 1936).

FIG 21 SEASONAL VARIATION IN EGG PRICES, ENGLAND AND WALES, 1930-37



SOURCE: Data of Table 7

hence adding to 1,200. The sum of the means is found to be 1,197·25. Each mean is adjusted by multiplying by $\frac{1.200}{1197{\cdot}25}$ = 1·0023, bringing the total to 1,200.

The estimate of seasonal variation is now in ratio and not in deviation form. Each month's figure appears as a percentage of the trend or norm. Thus January is 102·3 per cent or 2·3 per cent seasonally above the norm; February is 94·8 per cent or 5·2 per cent seasonally below the norm; and so on.

8.5 *Elimination of Seasonal Variation.* The effect of seasonal variation can now be eliminated from the original series. When the season is computed as a *deviation*, it is eliminated by the *subtraction* of the seasonal figure from the original value each month. When it is computed as a *ratio*, it is eliminated by *division* of the seasonal figure into the original value. In the example of 8·4, the seasonal variation in January is a positive deviation, or a ratio greater than 100 per cent. To eliminate it, each January price is reduced by a positive amount, or it is

divided by a number greater than one. In either case, the January price is reduced to allow for the fact that it is seasonally high in the month. The opposite holds for February prices which are seasonally low.

Elimination of seasonal movements is often needed in practice to give a month by month comparison of all variations in a time series (whether trend or residual) apart from the normal seasonal changes. Consider the monthly egg prices shown below for 1938 (data of Table 7). The general trend and any irregular movements are alike hidden by the presence of a strong seasonal variation. For example, the December price (212 d.) is below that in November (261 d.). But is the fall larger or smaller than the normal seasonal decline in price between these months? In other words, is the price up or down in December apart from the seasonal factor? The answer is to be sought after the seasonal variation is eliminated.

The seasonal variation is estimated (8.4 above) in deviation and in ratio form for the period 1934–7. It is now assumed that this seasonal variation continues unchanged into 1938. The series with seasonal factors eliminated is then obtained by either of the two alternative processes shown:

	Month	Price (pence)	Season (dev.)	Season elim. (pence)	Season (%)	Season elim. (pence)
		(1)	(2)	(3)	(4)	(5)
1938	Jan.	194	+ 3	191	102·3	190
	Feb.	174	— 8	182	95·0	183
	Mar.	124	— 51	175	68·4	181
	April	124	— 63	187	61·4	202
	May	136	— 57	193	64·9	210
	June	146	— 40	186	75·9	192
	July	184	— 17	201	89·8	205
	Aug.	201	+ 14	187	109·1	184
	Sept.	221	+ 17	204	110·3	200
	Oct.	239	+ 55	184	133·8	179
	Nov.	261	+ 86	175	152·3	171
	Dec.	212	+ 61	151	136·8	155

Note.—Col. (3) = Col. (1) — Col. (2); Col. (5) = $\frac{\text{Col. (1)}}{\text{Col. (4)}} \times 100$

With the seasonal effect eliminated, egg prices were high in the summer and fell rapidly in the last three months of the year.

8.6 *Components of a Time Series.* Each item in a short series of monthly figures has been split into three components—the short-run trend, the normal seasonal movement and the residual or irregular variations from month to month. The trend includes what is to be regarded as an oscillation from the point of view of the longer-run. The analysis applies to all or any part of the period used in estimating the trend and seasonal variation. The three components cumulate by *addition* when the method is to take *deviations* from the trend and by *multiplication* when *ratios* to trend are used.

The series of egg prices in the period 1934–7 (data of Table 7) illustrate the analysis. The computations are given in 8.4, and the complete analysis for the year 1937 is:

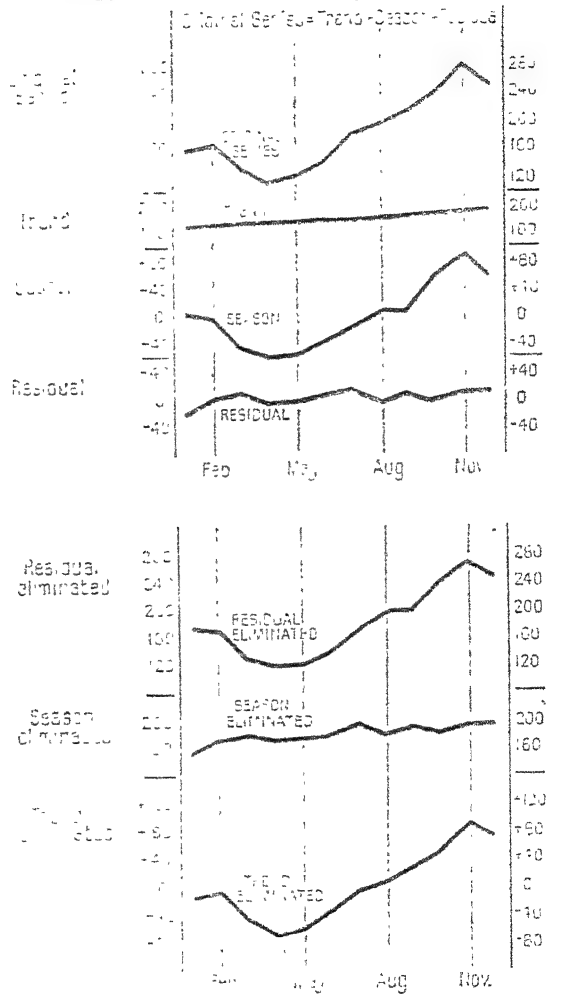
Month	Price (pence)	Analysis by deviations Trend + Season + Resid.					Analysis by ratios Trend × Season × Resid.				
	(1)	(2)		(3)		(4)	(5)		(6)		(7)
1937 Jan.	155	172	+	3	−	20	172	×	1·023	×	0·881
Feb.	164	173	−	8	−	1	173	×	0·950	×	0·998
Mar.	130	175	−	51	+	6	175	×	0·684	×	1·086
April	108	176	−	63	−	5	176	×	0·614	×	0·999
May	116	177	−	57	−	4	177	×	0·649	×	1·010
June	141	179	−	40	+	2	179	×	0·759	×	1·038
July	178	182	−	17	+	13	182	×	0·898	×	1·089
Aug.	191	184	+	14	−	7	184	×	1·091	×	0·951
Sept.	212	184	+	17	+	11	184	×	1·103	×	1·044
Oct.	238	185	+	55	−	2	185	×	1·338	×	0·961
Nov.	279	186	+	86	+	7	186	×	1·523	×	0·985
Dec.	255	187	+	61	+	7	187	×	1·368	×	0·997

Note.—Col. (4) = Col. (1) − Col. (2) − Col. (3);

$$\text{Col. (7)} = \frac{\text{Col. (1)}}{\text{Col. (5)} \times \text{Col. (6)}}$$

Each price can thus be split up into three additive parts as illustrated in Fig. 22A. Various combinations are possible, and some are shown in Fig. 22B. For example, the first series in Fig. 22B is the trend plus the season, i.e., the original series with residual variations eliminated. The alternative is to split each price into three multiplicative parts, with very similar results.

FIG.22 VARIATION IN EGG PRICES. ENGLAND AND WALES, 1937.



SOURCE - Data 6, Table 7

8.7 *Linear Trends.* The estimation of trend by moving averages has disadvantages. It is known that a moving average may introduce in the trend a spurious oscillation which is not really present in the time series.[1] This will result in a spurious oscillation in the deviations from trend and, hence, a distortion of the estimate of seasonal variation. There are other methods of proceeding not subject to this objection.

In one such method, a decision is made first on the *form* of the trend of a time series, e.g., that the trend is linear and represented by a straight line on the graph of the series. The problem then is to determine that particular straight line which "fits" best to the points of the series, a problem of the same kind as that of "fitting" a regression line to means of arrays. The solution generally adopted is to fit the linear trend to a time series by "least squares," to take it as the regression line of the variable on time as determined by the formula of 7.6 above.[2] A linear trend, so obtained, has the sum of squares of deviations of the values of the times series from the trend as small as possible. The straight line *AB* in Fig. 19 is a linear trend fitted to the time series of earnings in the period 1880–1914 (data of Table 5, Appendix I).

It is appropriate to fit a linear trend to the monthly series of egg prices, 1934–7 (data of Table 7, Appendix I), as an alternative to the moving average trend. An estimate of seasonal variation, slightly different from that found in 8·4 above, is then obtained by eliminating the linear trend. Simplifications and variations of this method of estimating seasonal variations, with correction for linear trend, are often used in practice, notably the method of link relatives as commonly employed in the U.S.[3]

8.8 *Correlation of Time Series.* Correlation is a statistical concept which is neutral as regards causal relations. The extent of correlation between two variables can be measured, but this does not determine which of the variables "causes" the other. Indeed, there may be no relation of cause and effect

[1] *See* Kendall, Appendix II, Ref. (24), Vol. II. Chap. 29.

[2] The practical computations are described in standard text-books, e.g., Croxton and Cowden, Appendix II, Ref. (6).

[3] *See* Croxton and Cowden, op. cit., Chap. XVII.

at all, the correlation showing only that outside factors influence both variables. This warning is particularly needed when we deal with time series.

Take the annual series of index numbers of cost of living (X) and of earnings (Y) in 1884–1900 (Appendix I, Table 5). There are seventeen pairs of values of X and Y, one pair for each year. The following figures can then be obtained:

$$\text{Mean } X = 88{\cdot}24 \qquad \text{Mean } Y = 81{\cdot}35$$
$$\text{Var } X = 9{\cdot}58 \quad \text{Var } Y = 34{\cdot}94 \quad \text{Cov } (XY) = -5{\cdot}48$$

So: correlation coefficient =

$$\frac{\text{Cov } (XY)}{\sqrt{\text{Var } X}\ \sqrt{\text{Var } Y}} = -\frac{5{\cdot}48}{\sqrt{9{\cdot}58 \times 34{\cdot}94}} = -0{\cdot}30$$

This is a surprising result. The negative correlation means that earnings tend to be high when the cost of living is low and conversely—not a very intelligible conclusion. An inspection of Fig. 19 helps to explain the result. There is a downward trend in the cost of living and an upward trend in earnings in the period; the negative correlation is an expression of these opposed trends. If there are only small variations about linear trends, the correlation would be nearly -1.

Our first point is that trends in time series contribute to, and often dominate, the correlation between them. Such correlation does not often interest us. In our example, earnings rose in 1884–1900 because of increasing productivity and other factors, *not* because the cost of living was falling. The negative correlation in the period does not help us to see whether earnings are affected by the cost of living. Even more "nonsensical" correlations can be obtained; for example, there is a strong positive correlation between the birth rate and the number of storks in Sweden—since each has been declining for various reasons. Such a correlation is a statistical fact; it just happens to be not very helpful.

To return to our example, we note that the correlation is no larger than $-0{\cdot}3$ though the trends are marked and nearly linear. This suggests that there is an opposite correlation between variations from each trend, and this may be more interesting. The next stage is to write deviations from trend, the plus and minus figures in the table of 8.3 above. Writing

x for the cost of living deviation and y for the earnings deviation in any year, we find for the period of seventeen years:

$$\text{Mean } x = -0{\cdot}771 \qquad \text{Mean } y = -0{\cdot}147$$

$$\text{Var } x = 5{\cdot}090 \qquad \text{Var } y = 5{\cdot}816 \qquad \text{Cov } (xy) = 4{\cdot}286$$

So: correlation coefficient =

$$\frac{\text{Cov } (xy)}{\sqrt{\text{Var } x \times \text{Var } y}} = \frac{4{\cdot}286}{\sqrt{5{\cdot}090 \times 5{\cdot}816}} = 0{\cdot}79$$

We do indeed find a strong positive correlation between variations from trend; if the cost of living is above trend then earnings tend to be above trend also. We can write the relation as a regression of earnings on cost of living (each as deviations from trend):

$$y - (-0{\cdot}147) = \frac{4{\cdot}286}{5{\cdot}090}\left(x - (-0{\cdot}771)\right)$$

i.e.,

$$y = 0{\cdot}842\ x + 0{\cdot}502$$

On the average for 1884-1900, if the cost of living index rises one point above trend, then the earnings index will be 1·3 points above trend; if the former index falls one point below trend, then the latter index will be 0·3 points below trend.

The correlation now measured arises largely because of oscillations in the two series which are connected with cycles of business activity (see Fig. 20). It would be desirable, though scarcely possible with our rough data, to push the analysis a stage further, i.e., to eliminate the regular oscillation in each series and then to correlate the residual deviations.

Our conclusion is that we must proceed warily in correlating time series since any correlation found may be due to trends or cycles in the series. If such correlations are not relevant, we must eliminate trends and/or cycles and correlate residual deviations. In any case, it is not easy to judge the significance of any result—which will depend on what kind of trend or cycle is fitted to the series. There is, however, an alternative to the stage-by-stage approach described so far. This is an extension of correlation and regression analysis to handle many variables at once, a method of great service with economic data. But this method cannot be dealt with here.

Much economic data are given as time series available, on a comparable basis, only for short periods. There are generally many variables at work and with high correlations between

them. One of the basic difficulties of empirical economics is that, while we are interested in complex relations between many variables, we have to work with very few observations. We cannot experiment and discover what a variable does under various conditions; we know only what its value was in a given month or year. The problem is a complicated one of regression and variance—of finding the extent to which one variable can be "explained" by others or a total variance broken down into "explained" and "unexplained" parts. We have been able to give here only a brief introduction to the analysis of this problem.

CHAPTER IX

SAMPLING AND SIGNIFICANCE

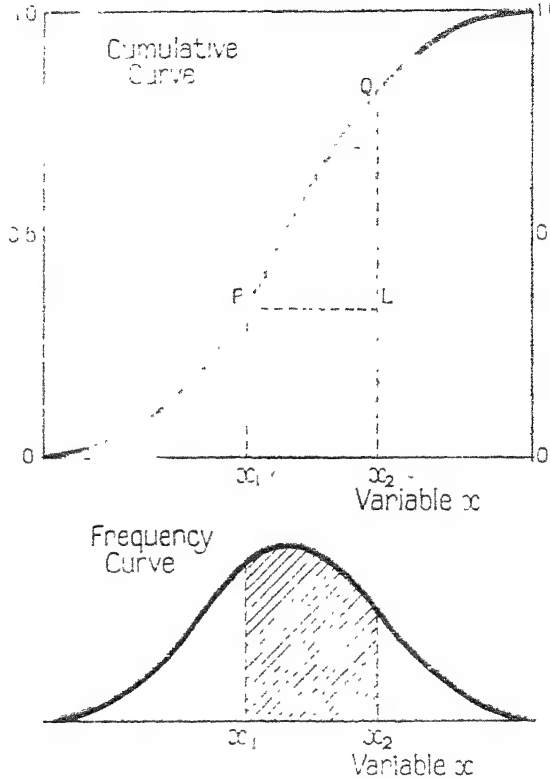
9.1 *The Normal Distribution.* There are several frequency distributions which appear often in the theory of statistics and which are much used in practice. The most common of these is the "normal" distribution, but there are many others such as the χ^2 (Chi-square), t and z distributions. The feature of these theoretical distributions is that they are *continuous*, represented by smooth curves of the kind shown in Fig. 15.[1]

An actual frequency distribution has only a limited number of observations of the variable; it is usually grouped in classes and shown as a block diagram as in Fig. 8 or 10. Theoretical distributions relate to a character x which is continuously variable and to a set of observations which can be taken as indefinitely large. Difficulties are avoided by writing *proportions* of the (indefinitely large) total as the frequencies of the distribution. In cumulative form, as illustrated in Fig. 23, the distribution shows the proportion of all cases with variable less than any specified x; the cumulative graph rises from 0 to 1 as x increases over its whole range. The frequency distribution and curve, in ordinary form, is derived from the cumulative graph. The *area* under the frequency curve from the left to any point x is the *height* of the cumulative curve at x, i.e., the proportion of cases with value less than x. So, in Fig. 23, the shaded area under the frequency curve is equal to the rise LQ in the cumulative curve, each being the proportion of cases with values of the variable between x_1 and x_2. This area is the probability of occurrence of values between x_1 and x_2 and, for this reason, the frequency distribution is often called a *probability distribution.* Attention must be concentrated on areas under a frequency curve, not on the heights of the curve.

The *normal distribution* is one example of a probability distribution, and it is defined by a particular mathematical formula. It is, however, not just one distribution, but a whole

[1]*See* Smith and Duncan, Appendix II, Ref. (19), pp. 100–14.

FIG. 23 HYPOTHETICAL CUMULATIVE AND FREQUENCY CURVES



set of distributions of the same general form. Two facts need to be known to fix it, the arithmetic mean $\bar{x}$ and the standard deviation σ of the variable. If $\bar{x}$ and σ are stated, the normal distribution is fixed completely; as different values are given, it takes different forms of the same general nature. The distribution can be represented by the *normal curve*; changing $\bar{x}$ merely moves the curve bodily to the left or right and changing σ merely stretches (or contracts) the curve horizontally. The

standard form of the normal distribution or curve has mean equal to zero and standard deviation equal to unity. All other normal curves can be obtained from it by moving bodily, and by stretching or contracting, horizontally. To get any normal distribution into standard form, the original variable x is switched to a new variable $z = \frac{x - \bar{x}}{\sigma}$. The distribution of z is that of x but in standard form; the switch is no more than a particular choice of units for measuring the variable. For example, if a distribution of families by rent (x shillings) has $\bar{x} = 5$ and $\sigma = 2$, then

x	1	2	3	4	5	6	7	8	9
$z = \frac{x - \bar{x}}{\sigma}$	-2	$-1\frac{1}{2}$	-1	$-\frac{1}{2}$	0	$\frac{1}{2}$	1	$1\frac{1}{2}$	2

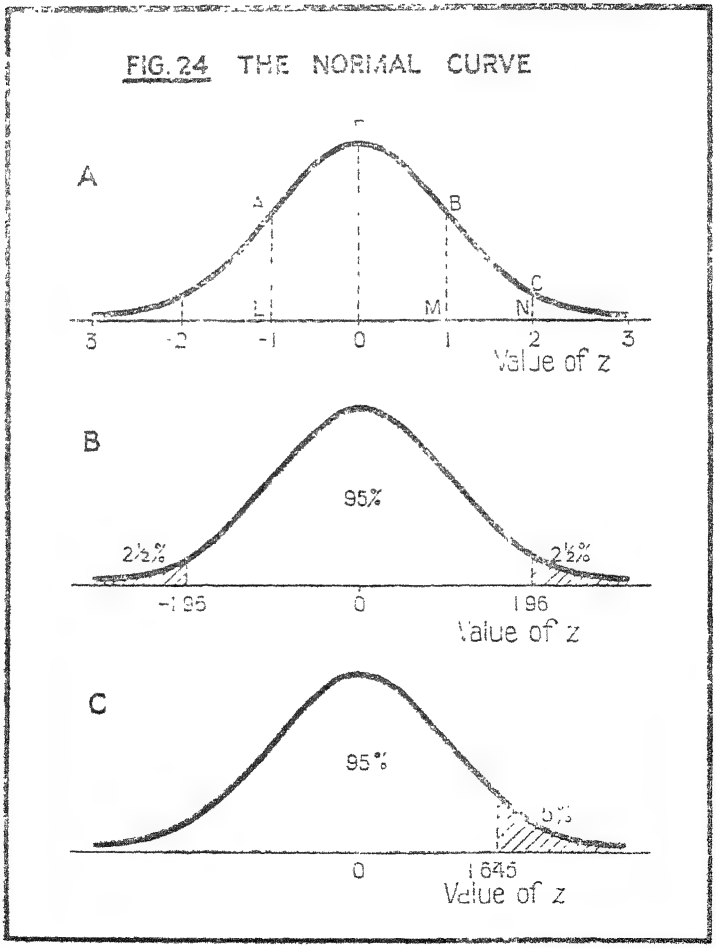
The original variable x is in shillings; the new variable is in units of the standard deviation (two shillings). If the normal curve is drawn with x markings on the horizontal scale, it is not in standard form; if with z markings on the scale, it is in standard form.

The general shape of the normal curve is the familiar "cocked-hat" form, symmetrical about the mean and stretching in long tails either way. The curve can be drawn from its mathematical formula, but this is not necessary in practice since tables have been prepared from the formula for use exactly like tables of logarithms. In condensed form:

Value of $z = \frac{x-\bar{x}}{\sigma}$	Normal curve		Value of $z = \frac{x-\bar{x}}{\sigma}$	Normal curve	
	Height at z	Area 0 to z		Height at z	Area 0 to z
(1)	(2)	(3)	(1)	(2)	(3)
0·0	0·3989	0·0000	1·2	0·1942	0·3849
0·2	0·3910	0·0793	1·4	0·1497	0·4192
0·4	0·3683	0·1554	1·6	0·1109	0·4452
0·6	0·3332	0·2258	1·8	0·0790	0·4641
0·8	0·2897	0·2881	2·0	0·0540	0·4773
1·0	0·2420	0·3413	2·5	0·0175	0·4938
			3·0	0·0044	0·4987

The total area under the curve is 1. The right-hand half of the curve is split into two parts by any z, the entry in col. (3) being the area from 0 to z, and the area beyond z being 0·5 *less* this entry.

Fig. 24A shows the normal curve in standard form plotted from col. (2) above. Col. (3) is sufficient to give any area



under the curve. For example, the proportion of all cases between $\bar{x} + \sigma$ and $\bar{x} + 2\sigma$ (i.e., z from 1 to 2) is

$$\begin{aligned}\text{Area } MNCB &= \text{Area } ONCP - OMBP \\ &= 0{\cdot}4773 - 0{\cdot}3413 = 0{\cdot}1360\end{aligned}$$

and the proportion between $\bar{x} - \sigma$ and $\bar{x} + 2\sigma$ (z from -1 to 2) is

$$\begin{aligned}\text{Area } LNCA &= \text{Area } LOPA + ONCP \\ &= \text{Area } OMBP + ONCP \\ &= 0{\cdot}3413 + 0{\cdot}4773 = 0{\cdot}8186\end{aligned}$$

One particular computation is to be of use later. The problem is to find a range of values of x or z containing 95 per cent of all items and the associated range with the other 5 per cent. This can be done in many ways, two of which are illustrated in Fig. 24. Fig. 24B shows the symmetrical range MN with 95 per cent of all items as represented by the unshaded area under the curve. The range below M and above N contains the other 5 per cent. Here, the area on ON is $\frac{1}{2}(0{\cdot}95) = 0{\cdot}475$ and N is seen from col. (3) above to occur at z a little less than 2. From more detailed tables, $z = 1{\cdot}96$. Hence:

In the normal distribution, 95 *per cent of all items occur in the range of* x *from* $\bar{x} - 1{\cdot}96\,\sigma$ *to* $\bar{x} + 1{\cdot}96\,\sigma$ *and* 5 *per cent of all items are more than* $1{\cdot}96\,\sigma$ *from* $\bar{x}$ *one way or the other.*

Fig. 24C shows a point L such that the range below L has 95 per cent and the range above L 5 per cent of all items. A detailed table of areas locates L at $z = 1{\cdot}645$. Hence, in the normal distribution, 95 per cent of all items occur below $\bar{x} + 1{\cdot}645\,\sigma$ and 5 per cent above this value of x.

Such ranges covering respectively 95 per cent, and 5 per cent of the total of cases are said to define the 5 *per cent level of significance.* Instead of 5 per cent, some other level of significance (say 1 per cent) can be specified and the corresponding ranges defined.

9.2 *Problems of Sampling.* It is not always practicable or even desirable to obtain data from a complete enumeration of items; information can often be derived quickly and cheaply, and with sufficient accuracy, from a sample of the total. The

totality of items is the *population* under investigation, and interest usually centres on a few characters or *parameters* of the population, for example the proportion of items with a given property or the mean value of a variable. A *sample* is drawn from the population and the measures of the corresponding characters in the sample serve as estimates of the population parameters; the mean of the variable in the sample, for example, is an estimate of the population mean. An estimate of a parameter from a sample is often called a *statistic*, a particular use of the singular of statistics. The whole problem of sampling is to use statistics from a sample to provide information about the population characters which are being investigated.

Generally, in economic and social inquiries, the population is an actual one with a definite (though often large) number of items. The rent data of Table 8, Appendix I, are from a sample of all agricultural holdings of five acres and over in England and Wales; the character sought is the mean rent per acre of all such holdings. In Appendix I, Table 10 or 14, the population is the group of all working-class families or households in the area surveyed; the parameter required may be the mean income of the whole group, the proportion of families with more than one person per room occupied, or some other character.

On some occasions, however, it is convenient to think up a hypothetical population. The most obvious cases come from games of chance. A pack of cards is an actual population of fifty-two items from which a sample "hand" of thirteen can be drawn. But suppose that a card is drawn, replaced, a second card drawn, replaced and so on. Then there is a hypothetical population of cards—say the first 1,000 cards we imagine drawn, or the indefinitely large number if cards are drawn without end. A sample of such a hypothetical population of cards consists of a set (say 100) of cards actually drawn. Similar hypothetical populations are useful concepts in experimental work (e.g., all temperature readings which might be taken as opposed to the few actually taken) and they have their applications also in the economic field.

There are many ways of drawing a sample of given size from a given population. The simplest is the *random sample*

where the items are selected at random, each member in the population having the same chance of being selected. The data of Tables 10 and 14 are obtained from samples which are (approximately) random. Another type of sample, increasingly used in practice, is the *stratified random sample* as employed in obtaining the rent data of Table 8, Appendix 1. The population is divided into groups or stratum according to some character (e.g., size of holding in Table 8) and random selections made within each stratum to make up the whole sample. Still another type is the *purposive sample* where a more deliberate selection of items is made according to certain criteria, as in the price data of Table 9, Appendix I, based on a "representative" sample of forty-five commodities.

In the following sections, the analysis is limited to random samples and to the case where the sample is large and the population very large. A sample will consist of n items drawn at random from a population of ν items, where n is large and ν so large that it can be taken as infinite. The theory and practice of sampling are much easier for large than for small samples. As a rough guide, a sample of upwards of 100 items can be regarded as large and a population of some thousands of items as infinite. A uniform notation will be adopted in which Roman letters indicate statistics from the sample, and the corresponding Greek letters the parameters of the population.

We can illustrate the problem with the example of a population of families for which we seek the proportion living under crowded conditions, defined as more than one person per room occupied. The population has ν families of which a proportion π are crowded; a random sample of n families is found to have a proportion p crowded. We wish to deduce something about π from knowledge of p. The problem can be handled in two stages, one theoretical and one practical.[1]

The theoretical question is, given a population with a definite parameter, to find the frequency distribution of occurrence of the corresponding sample statistic *in repeated samples of a given size* drawn from the population. The totality of these sample values makes up a distribution which is best put in probability or proportionate form. For example, there

[1]On these problems, *see* Smith and Duncan, op. cit., Chap. VIII.

is a distribution of sample proportions p when repeated samples of n families are taken from a population with proportion π of crowded families. The distribution tells us what percentage of the samples have proportions crowded between any two figures we care to name. The theoretical problem is a matter of mathematics which we can always take as solved. Our particular need is for the mean and standard deviation of the sampling distribution, i.e., the mean value assumed by the statistic in repeated samples and the dispersion over different samples. The standard deviation of a sampling distribution is so important that it is given a special name, the *standard error* of the character in sampling.

The practical problem reverses the process and attempts to argue back from one sample to the population. It can be approached in two ways:

(1) *Testing a given hypothesis.* We put up as a hypothesis a definite statement about the population parameter. We then test whether the actual sample could have arisen from such a population by the process of random sampling. For example, we may take $\pi = 0{\cdot}5$ in the population (50 per cent crowded families) as our hypothesis and test whether $p = 0{\cdot}4$ found in a sample (40 per cent crowded families) could arise by random sampling. The test of a hypothesis never gives a cut and dried answer of yes or no. Practically any p could arise from the given population. But some p's are more likely than others. The answer must then appear in terms of chances or probability. There are two types of error to be guarded against—the error of rejecting the hypothesis when it is true and the error of accepting it when it is false. These errors work in opposite directions since one error is increased in the process of reducing the other; the practical problem is one of compromise.

(2) *Estimating population parameters.* Here we make no hypothesis about the population, but we attempt to estimate the population parameter from the evidence of the sample. There are two things to do—first to get the best *single estimate* of the population parameter and then to specify the *range* around the estimate in which we can confidently assert the population parameter lies. In our example, we may estimate the population proportion of crowded families as p given by

the sample, but we must also attach an appropriate range of error. Our estimate usually appears in the form $p \pm e$, and we assert that we expect the population proportion to lie between $p - e$ and $p + e$. Again it is all a matter of "expectation," of chance or probability. The two types of error are still present —the error that the true population proportion in fact lies outside the range and the error of making the range so wide as to include too much to be of any practical use.

9.3 *Sampling for a Proportion.* A population of ν items (ν so large as to be taken as infinite) contains a proportion π of items with a given character. A large random sample of n items has a proportion p with the character. If repeated samples of n are taken, the values of p make up a frequency distribution which theory shows[1] is approximately normal in form with mean π and standard error $\sqrt{\frac{\pi(1-\pi)}{n}}$. Hence, 95 per cent of all values of p lie in the range

$$\pi \pm 1{\cdot}96 \sqrt{\frac{\pi(1-\pi)}{n}}$$

In only five samples out of every 100 does p lie outside this range.

It follows that the accuracy of the proportion found in a sample increases with the size n of the sample, not indeed in proportion to n but in proportion to $\sqrt{n}$. A sample must be increased four-fold to double its accuracy; a sample of 1,000 is twice (and not four times) as accurate as one of 250. For $\pi = 0{\cdot}5$, the value of $1{\cdot}96 \sqrt{\frac{\pi(1-\pi)}{n}}$ is 0·06 when $n = 250$, 0·03 when $n = 1{,}000$, and 0·01 when $n = 10{,}000$. With safety (at the 5 per cent risk) a sample of 250 will not quite give the proportion correct to the first decimal place, and a sample of more than 10,000 is needed to give the second decimal place correctly. On the other hand, it is a waste of resources to increase the size of the sample more than is needed to give

[1] *See* Smith and Duncan, op. cit., pp. 187–94. The sampling distribution is actually that known as the binomial.

the required accuracy; a larger sample than 10,000 is wasteful if the proportion is only required within 0·01.

In practice, π is not known, and only one value of p from a given sample is provided. We can proceed:

(1) *Testing the hypothesis that π has a definite value π_0.* In 95 per cent of all samples, the value of p lies in the range $\pi_0 \pm 1{\cdot}96\sqrt{\frac{\pi_0(1-\pi_0)}{n}}$. Suppose we *accept* the hypothesis if the given value of p lies in this range and *reject* it if p falls outside the range. Then we reject the hypothesis when true in only five cases out of 100. But we accept the hypothesis quite often when it is false. In general, we make the best compromise on this risk by taking the symmetrical range for p rather than some other range including 95 per cent of all cases.[1] The range specified is that for the 5 *per cent level of significance* since true hypotheses are rejected in only 5 per cent of all applications of the rule. Other levels can be taken, but the 5 per cent level is generally convenient in practice.

(2) *Estimating the unknown value of π.* The best single estimate of π is the value p given by the sample. For, if the population proportion is p, then the distribution of sample proportions in repeated samples is normal with mean p, i.e., the actual sample p is the most likely of all to occur. It is not easy to define the range in which we can expect π to lie, but it can be shown that, approximately, the range

$$p \pm 1{\cdot}96\sqrt{\frac{p(1-p)}{n}}$$

will include the population value π in 95 per cent of all occasions in which the rule is applied.[2] That is, in using this range for π, we run a 5 per cent risk of failing to include the correct but unknown π. Further, by taking the symmetrical range, rather than some other at the 5 per cent level of significance, we make the range as narrow as possible. The 5 per cent level is again adopted as convenient in practice.

[1] *See* Smith and Duncan, op. cit., pp. 198–201, for an extended account of this problem.

[2] *See* Smith and Duncan, op. cit., pp. 202–4.

From Appendix I, Table 14 (a sample of working-class families in Kensington, 1929–30), 233 of a total of 484 families are seen to have more than one person per room. The proportion crowded is 48 per cent — $p = \frac{233}{484} = 0{\cdot}48$. In the population census of 1931, the proportion of all families crowded is found to be $\pi = 0{\cdot}305$ in Kensington. Can it be asserted that the group of 484 *working-class* families is a random sample of *all* families in Kensington as regards this property? In repeated samples of 484 from a population with $\pi = 0{\cdot}305$, the value of p found will lie in the range

$$0{\cdot}305 \pm 1{\cdot}96\sqrt{\frac{0{\cdot}305 \times 0{\cdot}695}{484}} = 0{\cdot}305 \pm 0{\cdot}041$$

in 95 per cent of all cases. The range is from 0·264 to 0·346, and the actual value of p is 0·48. Hence, at the 5 per cent level of significance, the sample is *not* likely to arise from this particular population—the evidence is that working-class families are more crowded than all families in Kensington.

What, then, is our estimate of the proportion of all working-class families living in Kensington in crowded conditions? By the rule given, the estimate is

$$\pi = 0{\cdot}48 \pm 1{\cdot}96\sqrt{\frac{0{\cdot}48 \times 0{\cdot}52}{484}} = 0{\cdot}48 \pm 0{\cdot}045$$

We shall be wrong in only 5 per cent of all cases in saying that the proportion sought lies between 43½ and 52½ per cent. The best single estimate is 48 per cent, and there is a margin (at the 5 per cent level of significance) of 4½ per cent either way in this figure.

9.4 *The Difference between Proportions in Two Samples.* Large samples of n_1 and n_2 items respectively are taken independently from two populations. The first is found to have a proportion p_1 with a given character, and the second a proportion p_2, where p_1 is greater than p_2. Nothing is known about the proportions π_1 and π_2 with the character in the two populations. Can it be asserted that π_1 is greater than π_2 on the evidence?

One method of approach is to estimate π_1 and π_2 separately

as in 9.3 above. With the 5 per cent level of significance, each estimate has a range of error. If the two ranges do not overlap then it is safe (at the 5 per cent risk) to say that π_1 is greater than π_2. If the ranges do overlap then this cannot be said with safety. This test, however, is not an efficient one; it rejects many differences which should be taken as significant. A significant difference between π_1 and π_2 is rejected in the test if the ranges just overlap—but this depends on the chance that π_1 is at the lower end of its range *and* that π_2 is at the upper end of its range. This chance is much smaller than the separate chances of π_1 and π_2 being at the ends of their ranges. In fact the level of significance involved in the test is really smaller than 5 per cent.

The method is improved by reverting to the process of testing a particular hypothesis. We put up the hypothesis that *the two populations have the same proportion* π, and we put it up so that we can knock it down if possible. We then test whether each of the two samples could have come from the common population, say at the 5 per cent level of significance. If they could not, we reject the hypothesis and we assert that π_1 is significantly greater than π_2. If they could, then we have a possible hypothesis, and we cannot assert that there is any significant difference between π_1 and π_2 on the evidence of the two samples. The basic result for the test is that the difference $(p_1 - p_2)$ in repeated samples of n_1 and n_2 from a population with proportion π is distributed approximately according to the normal law with mean zero and variance equal to the sum of the variances of p_1 and p_2 separately. These latter variances are $\frac{\pi\,(1-\pi)}{n_1}$ and $\frac{\pi\,(1-\pi)}{n_2}$ respectively. Hence:

$$\text{Standard error of } (p_1 - p_2) = \sqrt{\pi\,(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

At the 5 per cent level of significance, we can say that π_1 is greater than π_2 if

$$(p_1 - p_2) \text{ is greater than } 1{\cdot}96\sqrt{\pi\,(1-\pi)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}.$$

To apply this test, we need to know π.

The best estimate we can make of π is the proportion found with the character in the two samples *pooled together*. This value is determined, substituted in the above formula, and the actual difference $(p_1 - p_2)$ found in the samples compared with its standard error.

In the sample data of Table 14, one group has $n_1 = 175$ families each occupying two rooms, and the proportion crowded is $p_1 = \frac{104}{175} = 0{\cdot}59$. Another group has $n_2 = 131$ families each with three rooms and a proportion crowded of $p_2 = \frac{59}{131} = 0{\cdot}45$. In the population as a whole, are there more crowded families amongst those with two rooms than amongst those with three rooms? As a first approach, we fix the 5 per cent level of significance, and we estimate π_1 (families with two rooms) and π_2 (families with three rooms) separately:

$$\pi_1 = 0{\cdot}59 \pm 1{\cdot}96\sqrt{\frac{0{\cdot}59 \times 0{\cdot}41}{175}} = 0{\cdot}59 \pm 0{\cdot}073$$

$$\pi_2 = 0{\cdot}45 \pm 1{\cdot}96\sqrt{\frac{0{\cdot}45 \times 0{\cdot}55}{131}} = 0{\cdot}45 \pm 0{\cdot}085$$

Hence π_1 ranges down to 0·52 and π_2 up to 0·535. The ranges just overlap and the test just fails to indicate a significant difference between π_1 and π_2.

In a second approach, the hypothesis is made that the two groups come from populations with the same proportion π of crowded families. The pooled samples give an estimate of

$$\pi = \frac{163}{306} = 0{\cdot}53.$$

The difference $(p_1 - p_2)$ has:

$$\text{Standard error} = \sqrt{0{\cdot}53 \times 0{\cdot}47\left(\frac{1}{175} + \frac{1}{131}\right)} = 0{\cdot}058$$

The actual difference in the two given groups is

$$p_1 - p_2 = 0{\cdot}14 = 2{\cdot}4 \text{ times the standard error.}$$

Since this multiple is greater than 1·96, the hypothesis put up can be rejected at the 5 per cent level of significance. Hence, with this 5 per cent risk, we can assert that the proportion of

crowded families is greater among those in two rooms than among those in three rooms.

9.5 *Sampling for a Mean.* A population of ν items (ν so large as to be taken as infinite) has a variable x following a normal distribution with mean $\bar{\xi}$ and standard deviation σ. It is assumed here that σ *is known and given.* In a large sample of n items, the variable x is found to have mean $\bar{x}$. Theory shows that, in repeated samples, the sampling distribution of $\bar{x}$ is normal with mean $\bar{\xi}$ and standard error $\frac{\sigma}{\sqrt{n}}$. Hence, in 95 per cent of all samples, the mean $\bar{x}$ found will lie in the range $\bar{\xi} \pm 1{\cdot}96 \frac{\sigma}{\sqrt{n}}$. Again, the accuracy of sampling increases as the size of sample increases, but only in proportion to $\sqrt{n}$.

The significance of any mean $\bar{x}$ can then be tested on lines similar to those adopted for a proportion above. Any specific hypothesis, e.g., that the population mean is $\bar{\xi}_0$, can be tested and the best estimate of the population mean can be made. At the 5 per cent level of significance, the estimate of the population mean is $\bar{x} \pm 1{\cdot}96 \frac{\sigma}{\sqrt{n}}$. Further, if two samples of n_1 and n_2 items are found to have means $\bar{x}_1$ and $\bar{x}_2$, where $\bar{x}_1$ is greater than $\bar{x}_2$, we can test whether the first population mean $\bar{\xi}_1$ is greater than the second $\bar{\xi}_2$. We set up the hypothesis that the population means are the same; the distribution of $(\bar{x}_1 - \bar{x}_2)$ in repeated samples is normal, as before, and with mean zero and standard error $\sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$. Hence, if the difference actually found is greater than $1{\cdot}96\ \sigma \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$, we reject the hypothesis, and we assert that, at the 5 per cent level of significance, the first population mean is greater than the second.

Table 14 shows that 10·5 shillings per week is the mean rent paid by 484 working-class families sampled in Kensington in 1929–30. No information is given about the distribution of

rents over the different families. For the purpose of the present analysis, it is assumed that the standard deviation of rents in the population (all families in the working-class in Kensington) is 3·5 shillings per week. This is a guess, but it is of the right order of magnitude. As a first problem, the hypothesis is tested that the mean rent in the population is ten shillings. In repeated samples of 484 families, the distribution of the sample mean $\bar{x}$ is normal with mean ten and standard error $\frac{3{\cdot}5}{\sqrt{484}}$. In 95 per cent of all samples, the mean $\bar{x}$ lies in the range

$$10 \pm 1{\cdot}96 \frac{3{\cdot}5}{\sqrt{484}} = 10 \pm 0{\cdot}31$$

i.e., between 9·7 and 10·3 shillings. The actual mean rent found is 10·5 shillings, outside the range. At the 5 per cent level of significance, the hypothesis that the population mean rent is ten shillings can be rejected.

The best estimate of the mean rent for all working-class families in Kensington can now be written:

$$\bar{\xi} = 10{\cdot}5 \pm 1{\cdot}96 \frac{3{\cdot}5}{\sqrt{484}} = 10{\cdot}5 \pm 0{\cdot}31$$

The best single estimate is 10·5 shillings and, at the 5 per cent level of significance, there is a margin of 0·3 shillings.

Table 14 also gives the mean rent paid by groups of families occupying one, two, three, . . . rooms. One sample (four rooms occupied) has $n_1 = 70$ families and mean rent $\bar{x}_1 = 14{\cdot}2$ shillings. Another sample (three rooms occupied) has $n_2 = 131$ families and mean rent $\bar{x}_2 = 12{\cdot}0$ shillings. Can it be asserted that the average rent paid by all Kensington working-class families occupying four rooms is higher than that paid by families with three rooms? Put up the hypothesis that the population mean rents are the same, and assume (again) that the standard deviation of rents is 3·5 shillings per week for families with three and four rooms. Then the difference $(\bar{x}_1 - \bar{x}_2)$ in samples has

$$\text{Standard error} = 3{\cdot}5 \sqrt{\frac{1}{70} + \frac{1}{131}} = 0{\cdot}52$$

The actual difference found in the two groups is

$\bar{x}_1 - \bar{x}_2 = 2{\cdot}2 = 4\frac{1}{4}$ times the standard error.

The multiple is much greater than 1·96, required by the 5 per cent level of significance. The hypothesis is rejected and the difference in the mean rents paid by families occupying three and four rooms is significant.

9.6 *Further Analysis of Sampling for a Mean.* The methods of 9.5 are subject to the serious limitation that the population σ must be known. The problem becomes more difficult when, as is usual, σ is unknown. An estimate of σ must be made in order to write the standard error of the mean in samples. The standard deviation of the variable in the given sample would seem to be the best estimate of σ, but this is not quite correct. It can be shown that the best estimate of σ is $\sqrt{\frac{n}{n-1}}$ times the sample standard deviation, i.e., it is s where

$$s^2 = \frac{n}{n-1} \times \frac{1}{n} \Sigma (x - \bar{x})^2 = \frac{1}{n-1} \Sigma (x - \bar{x})^2$$

For a variable x in a sample of n items, the standard deviation is $\sqrt{\frac{1}{n} \Sigma (x - \bar{x})^2}$. But, if an estimate of the population standard deviation is needed, the slightly different figure $s = \sqrt{\frac{1}{n-1} \Sigma (x - \bar{x})^2}$ is computed instead.[1]

The further complication now arises that, with the substitution of s for σ in the standard error of the mean, the sampling distribution of the mean is no longer normal. It follows a distribution, known as the t-distribution, which differs considerably from the normal distribution when the size n of the sample is small, but which is approximately normal when n is large.[2] This still assumes that the distribution of the variable x in the population itself is normal. The results must be modified further when the population distribution takes some other form, e.g., if it is J- or U-shaped.

To sum up, we can say that the methods of 9.5 apply *approximately* when s obtained from the sample is substituted

[1] *See* Smith and Duncan, op. cit., pp. 181–2 and 290–4.
[2] *See* Smith and Duncan, op. cit., pp. 109–11 and 241.

for σ, if we are testing the mean $\bar{x}$ of a variable in a *large sample*, and if we can take the distribution of the variable in the population as *nearly normal* in form. If the sample is small, or even only moderately large, we should test the sample mean with the *t*-distribution and not by the methods of 9.5 based on the normal distribution. Still more elaborate methods are needed if the population distribution is definitely not normal. The methods of 9.5 do not apply, for example, if a sample of coal mines is taken to test a variable connected with size of mine; the distribution of mines by size is J-shaped (Appendix I, Table 13).

The analysis of variance provides a different approach to the problem of comparing means in different samples.[1] This approach has the double advantage of applying to small samples and of comparing several means together. The variance of the means among themselves is compared with the variance of the original values of the variable about each separate mean, very much on the lines of 7.7 above. This approach is connected with the notion of correlation and helps to determine whether the progression of mean values from one group to another is the result of correlation between the variable and the character which describes the groups. In Table 14 of Appendix I, for example, mean rent paid is seen to increase among families occupying an increasing number of rooms. An analysis of variance would show whether the means are significantly different and hence whether there is correlation between rent and size of dwelling. For such an analysis, full information, not given in Table 14, is needed on rents paid by all families in each group.

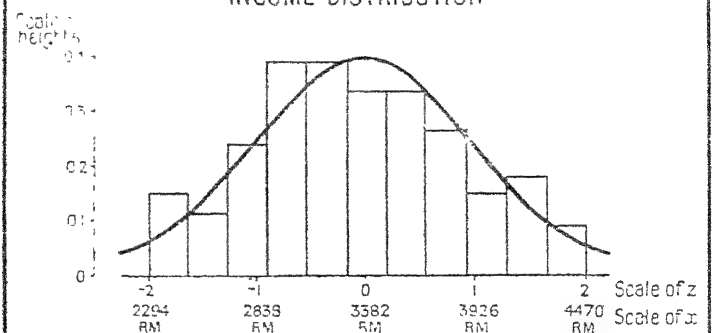
9.7 *Curve Fitting.* The normal distribution is of very wide application. We have noted its use in sampling problems. It is also a good "fit" to the distributions of many variables found in practice. Speaking broadly, we can say that a variable follows the normal law when its value arises from a large number of factors each contributing independently a small amount to the value. For this reason the normal distribution is sometimes called the "law of error" or the "law of large

[1] *See* Tippett, Appendix II, Ref. (17), Chap. VI, and particularly pp. 136–8.

numbers." It is the distribution to be expected in experimental observations (e.g.) of temperature. It is equally the distribution to be expected in an economic variable which is governed by many small factors.

The method of fitting a normal to an actual distribution is simple; the mean and standard deviation of the normal distribution, which are at choice, are taken equal to the mean and standard deviation of the actual distribution. The result of such a fitting is shown in Fig. 25, which relates to the income

FIG 25 NORMAL CURVE FITTED TO INCOME DISTRIBUTION



SOURCE: Data of Table 10B

distribution of the ninety families of Appendix I, Table 10B. The distribution is expressed in proportions of the total of ninety families. The fit is apparently quite good.

However, we should determine more precisely how good the fit is. This can be regarded as a problem in sampling. We put up the hypothesis that the actual distribution is a random sample of a population in which the variable follows the normal law. If we can accept this hypothesis at a certain level of significance, then we say that the fit is good. A test designed for this purpose is that known as the χ^2 (Chi-square) Test.[1]

9.8 *Conclusion.* A sample has been taken here as a substitute for complete enumeration, adding a second type of error—arising from taking one sample rather than another—

[1]*See* Smith and Duncan, op. cit., pp. 137–42.

to the errors inherent in all statistical figures. This distinction, however, disappears on closer analysis. The time series of employment (Appendix I, Table 6) is obtained by complete enumeration of insured workers and of those unemployed. The frequency distribution of mines by size (Appendix I, Table 13) is similarly derived from data on all mines. There are many and various factors at work making employment what it is or affecting the size of a mine. Some of them are of major import, e.g., the influence of the "trade cycle" in one case and of geological factors in the other. Others have a minor or erratic effect, including sheer inaccuracies in recording. Once the effect of major factors is isolated and eliminated, the residual factors left can be treated as a problem in sampling. A particular residual value actually recorded may be regarded as the result of a particular sample of many small, and often accidental, influences at work. The various residuals may be expected to follow the normal law.

For example, the methods of deriving trends and cyclical fluctuations, and the technique of correlation, serve to isolate the major factors in time series. But there always remain certain residual fluctuations, and here the sampling approach becomes appropriate. Each value of a time series is one record which may be viewed as a particular sample of all the different hypothetical values which could have arisen. Similarly, a frequency distribution is a set of actual records of a variable. It may be regarded as one sample from a hypothetical population in which the variable is distributed according to some "regular" law, a law which is blurred by random influences in the actual recording. This is what lies behind and justifies the method of curve fitting.

So, by the use of hypothetical populations and the identification of an actual record as one out of many possible outcomes, the concepts of sampling and significance can be applied over a very wide field. We can end our development of statistical methods here with the thought that sampling is not just a specialized technique, but rather the central problem of statistics itself. The methods of sampling and significance are needed in all statistical methods, which can be said, broadly, to concern the testing of hypotheses about the variation of characters under the influence of various factors.

APPENDIX I

ILLUSTRATIVE TABLES

THE following tables serve a double purpose; they illustrate various points discussed in the first chapter, and they provide suitable examples of the practical application of statistical methods developed in subsequent chapters. Though some preliminary conclusions are added to the tables, these are not complete analyses of the data and they do no more, generally, than raise points which need further investigation.

TABLE 1

NET NATIONAL INCOME AND EXPENDITURE, U.K. (£ millions)

INCOME			EXPENDITURE		
	1938	1946		1938	1946
	(1)	(2)		(3)	(4)
Rent	380	386	Personal consumption	3,668	5,420
Interest and profits (*a*)	1,368	2,370	Government current expenditure (*c*)	765	2,261
Salaries	1,110	1,675			
Wages	1,735	3,020	Net investment		
Pay and allowances of Armed Forces (*b*)	78	523	At home	308	693
			Abroad	− 70	− 400
Net National Income	4,671	7,974	Net National Expend.	4,671	7,974

From: National Income and Expenditure of the U.K., 1938 *to* 1946 (Cmd. 7099, 1947).

Notes: Estimates made in the Central Statistical Office using Inland Revenue and other data furnished by government departments. Net national income is at factor cost (i.e., adjusted for indirect taxes and subsidies) and excludes depreciation allowances and transfer payments. Consumption, government expenditure and investment are similarly at factor cost. The primary estimate is of net national income; expenditure is separately estimated except for a residual item (net investment at home) inserted to give net national expenditure equal to net national income.

(*a*) Including farming profits and professional earnings.

(*b*) In cash and kind, serving members of the Forces only.

(*c*) On goods and services of all kinds.

Object of table: To make a broad survey of the national economy in money terms, showing how national income is earned and how it is spent.

Some conclusions: As compared with the pre-war year of 1938, incomes from rent and salaries declined in 1946 relative to wages and profits. Government expenditures and net investment at home were much higher. Personal consumption was nearly 50 per cent up in value, but it can be shown that the over-all volume of consumption in 1946 was much the same as in 1938. This was made possible by borrowing from abroad and by sale of assets to foreigners, together making up the negative figure for net investment abroad.

TABLE 2

Distribution of Total Working Population, Great Britain

	June 1939	June 1945	June, 1946		
	000's	000's	000's	% of June, 1939	% of June, 1945
	(1)	(2)	(3)	(4)	(5)
Armed Forces	480	5,090	2,032	423·3	39·9
Government (*a*)	1,465	2,030	2,099	143·3	103·4
Agriculture (*b*)	950	1,041	1,078	113·5	103·6
Mining	873	799	806	92·3	100·9
Building	1,310	722	1,184	90·4	164·0
Manufacturing:					
Equipment for Forces	1,270	3,830	715	56·3	18·7
Export	990	410	1,310	132·3	319·5
Home market	4,555	2,580	4,562	100·2	176·8
Other industries (*c*)	6,587	5,004	5,661	85·9	113·1
Not in employment (*d*)	1,270	143	1,076	84·7	752·4
Total working population	19,750	21,649	20,523	103·9	94·8

From: Ministry of Labour *Gazette* and *Monthly Digest of Statistics.*

Notes: Estimates based on mid-year count of insured workers, monthly returns of employment from firms, and other data available to the Ministry of Labour. Males aged 14—64 and females aged 14—59; employers, self-employed or in paid employment (except private domestic servants), together with insured persons registered as unemployed and ex-Service men and women not yet taken up employment. Those sick, on holiday or otherwise absent from work included as employed if maintained on employers' books. Women in part-time paid employment included, each counted as half a full-time worker.

(*a*) National and local government, Civil Defence, N.F.S., and police, but excluding trading services.

(*b*) Including horticulture and fishing.

(*c*) Transport, shipping, public utilities, distribution, and services.

(*d*) Including ex-Service men and women not employed since discharge, numbering 700,000 in June, 1946, of whom 35,000 registered as unemployed.

Object of table: To examine the extent of reconversion in the first post-war year, 1945–6, and to compare with the immediate pre-war period.

Some conclusions: Considerable progress towards achievement of post-war export goal (75 per cent above 1938); a return to the pre-war level in manufacturing for the home market; little increase in employment in coal-mining; building and many service trades far short of pre-war levels; unemployment at low levels, apart from ex-Service men and women not employed since discharge.

TABLE 3

Exports of Beverages and Cocoa Preparations, U.K.

A. *Quantities and Average Values.*

Commodity	Unit	000 units			£ per unit		
		1935	1938	1946	1935	1938	1946
		(1)	(2)	(3)	(4)	(5)	(6)
Spirits (*a*)	Proof gal.	6,396	9,117	6,536	1·194	1·246	1·803
Beer	Std. barrel	224·1	271·1	150·1	4·89	4·22	14·34
Fruit juice and table waters	gal.	440·9	587·1	402·4	0·370	0·381	0·548
Cocoa preparations:							
Containing sugar	cwt.	115·6	104·3	212·9	4·62	5·25	7·24
Not containing sugar	cwt.	235·9	312·7	388·9	0·620	0·847	1·377

B. *Valuation of Exports* (£000).

Commodity	At 1935 av. values		At 1938 av. values			At 1946 av. values	
	1935	1938	1935	1938	1946	1938	1946
	(7)	(8)	(9)	(10)	(11)	(12)	(13
Spirits (*a*)	7,634	10,885	7,969	11,362	8,143	16,437	11,782
Beer	1,095	1,326	946	1,143	633	3,887	2,153
Fruit juice and table waters	163	217	168	223	153	322	221
Cocoa preparations:							
Containing sugar	534	482	607	548	1,118	755	1,542
Not containing sugar	146	194	200	265	329	431	536
All other items (*b*)	193	274	199	284	624	457	976
Total	9,766	13,378	10,089	13,824	11,000	22,289	17,209

From: Annual Statement of Trade of the U.K. and *Accounts Relating to the Trade and Navigation of the U.K.*

Notes: Recorded exports (Class I, Group G) of produce and manufactures of the U.K., by quantity, cols. (1), (2) and (3), and by value, cols. (7), (10) and (13), as reported on returns received by the Board of Trade in the year. Average values, cols. (4), (5) and (6), obtained by division of value by quantity and used in the valuations of cols. (8), (9), (11) and (12). Spirits are home-made only; fruit juice and cocoa preparations exclude items containing spirits; cocoa preparations containing sugar are mainly chocolate.

(*a*) Ethyl alcohol for industrial purposes included in 1935 and 1938, excluded in 1946.
(*b*) Recorded values in cols. (7), (10) and (13); other values obtained on assumption that price change equals that in five specified items together.

Object of table: To show changes in the value of a group of exports between 1935, 1938 and 1946, and to analyse them into the part due to changing prices and the part due to the changing volume of trade.

Some conclusions: Changes in trade in individual commodities are varied; in general, prices in the group changed little from 1935 to 1938, and then increased by nearly 60 per cent to 1946; volume of exports increased by more than one-third from 1935 to 1938, but returned to little more than the 1935 level in 1946 (*see* 6.4).

TABLE 4

Construction of an Index Number of Retail Food Prices, U.K.

Item	Unit	(a) Weights	July, 1914		1st Sept., 1939		Products		
			Purchases (Units)	Prices (d)	Prices (d)	% of July 1914	(1) x (5)	(2) x (3)	(2) x (4)
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Beef,									
British	lb.	24	1·95	8½	10⁷	131	3,144	16 09	21·18
Imported	lb.	24	2·07	6	7½	119	2,856	16 02	19 01
Mutton									
British	lb.	12	0·95	8½	11¼	135	1,620	8 07	10·93
Imported	lb.	12	1·51	5⁷	7⁵	133	1,596	8·12	10·75
Bacon	lb.	19	1·11	11½	15	133	2,527	12·49	16·65
Fish (*b*)	..	9	...	...	...	216	1,944	6·00	12·96
Flour	6 lb.	20	1·50	9	11½	128	2,560	13·50	17·25
Bread	4 lb.	50	5 9	5¾	8¼	143	7,150	33·92	48·68
Tea	lb.	22	0·81	18½	28	153	3,366	14 78	22·68
Sugar	lb.	19	6 5	2	3	150	2,850	13 00	19·50
Milk	quart	25	4·8	3½	6⅝	193	4,825	16 80	32·40
Butter	lb.	41	1·95	14⁷	15⁵	110	4,510	28 03	30·93
Cheese	lb.	10	0 80	8¾	10	114	1,140	7·00	8·00
Margarine	lb.	10	0·93	7	6½	93	930	6·51	6·05
Eggs	each	19	10·6	1¼	2	160	3,040	13·25	21·20
Potatoes	7 lb.	18	2 53	4¾	6½	137	2,466	12·02	16·46
Total		334					46,524	225·60	314·63

From: Ministry of Labour *Gazette.*

Notes: Purchases based on budget of 1914 derived from inquiry made in 1904 (Cd. 2337) with minor modifications to apply to 1914. Price quotations (from representative shops with working-class trade, averaged for large and small towns) from returns made monthly by employment exchanges of Ministry of Labour. The quality of the items should remain unchanged, but this is not always possible in practice (e.g., tea).

(*a*) Approximately proportional to expenditure in 1914 = col. (7).

(*b*) Purchases and price quotations not given; cols. (7) and (8) derived from assumed expenditure of 6d. in 1914 and 116 per cent increase in price.

Object of table: To compare food prices paid by working-class families in 1939 with those of 1914, and to measure the change in the general level of food prices.

Some conclusions: Changes in food prices from July, 1914 to 1939, vary from a fall of 7 per cent (margarine) to a rise of 116 per cent (fish); change in general level of prices paid by the working-class families is shown by the index number of 139 at 1st September, 1939 (July, 1914 = 100). This is col. (6) divided by col. (1) equals col. (8) divided by col. (7) (*see* 6.3).

TABLE 5

Earnings and Cost of Living, U.K., 1880–1938
(Index numbers, 1914 = 100)

Year	Cost of living	Full-time earnings	Real earnings	Year	Cost of living	Full-time earnings	Real earnings
	(1)	(2)	(3)		(1)	(2)	(3)
1880	105	72	69	1910	96	94	98
1881	103	72	71	1911	97	95	97
1882	102	75	73	1912	100	98	97
1883	102	75	73	1913	102	99	97
1884	97	75	77	1914	100	100	100
1885	91	73	81	1915	123	...	...
1886	89	72	81	1916	146	...	...
1887	88	73	84	1917	176	...	...
1888	88	75	86	1918	203	...	...
1889	89	80	90	1919	215	...	...
1890	89	83	93	1920	249	...	...
1891	89	83	92	1921	226	...	...
1892	90	83	92	1922	183	...	...
1893	89	83	94	1923	174	...	...
1894	85	83	98	1924	175	194	111
1895	83	83	100	1925	176	196	112
1896	83	83	100	1926	172	195	113
1897	85	84	98	1927	167½	196	117
1898	88	87	99	1928	166	194	117
1899	86	89	104	1929	164	193	118
1900	91	94	103	1930	158	191	121
1901	90	93	102	1931	147½	189	128
1902	90	91	101	1932	144	185	129
1903	91	91	99	1933	140	183	131
1904	92	89	97	1934	141	186	132
1905	92	89	97	1935	143	191	132
1906	93	91	98	1936	147	197	134
1907	95	96	101	1937	154	207[1]	134
1908	93	94	101	1938	156	212[1]	136
1909	94	94	100				

From: Bowley, *Wages and Income Since* 1860 (1937), Table VII, extended to 1938.

Notes: Cost of living series is official index (1915–38) with earlier figure (1880–1914) estimated by Bowley as far as possible on a comparable basis. Full-time earnings estimated by Bowley as "average earnings for a normal week of all wage-earners in the U.K., the changes in the relative numbers in different occupations and industries being taken into account." Real earnings obtained by division by cost of living index.

[1]Rough estimates extending Bowley's series.

...Not available.

Object of table: To show long-period changes in money and real earnings of the working class from 1880, the first year for which adequate data are available.

Some conclusions: Retail prices and money earnings tend to follow each other quite closely except for an upward secular movement in earnings. The latter, shown by changes in real earnings, is associated with increasing productivity of labour. The growth in real earnings, rapid until 1895, then flattened off until 1914, and was resumed in the period between the two world wars (*see* 8.2 and Fig. 21).

TABLE 6

DEPOSITS AND OTHER ACCOUNTS OF CLEARING BANKS AND EMPLOYMENT, GREAT BRITAIN, 1921–46

Year	Clearing Banks			Employment (July) %
	Deposits £ mn.	Advances £ mn.	Investments £ mn.	
	(1)	(2)	(3)	(4)
1921	1,768	815	309	82·6
1922	1,727	732	372	87·1
1923	1,631	744	338	88·6
1924	1,632	791	324	90·4
1925	1,623	839	270	89·2
1926	1,626	876	249	85·8
1927	1,675	888	238	90·9
1928	1,729	933	239	88·7
1929	1,762	974	242	90·4
1930	1,763	948	243	83·5
1931	1,723	904	285	78·2
1932	1,752	830	332	77·4
1933	1,914	746	519	80·7
1934	1,842	740	543	83·4
1935	1,961	755	598	85·0
1936	2,104	825	598	87·8
1937	2,172	910	607	90·1
1938	2,161	930	593	(*a*)86·9 (*b*)89·0
1939	2,129	943	564	92·4
1940	2,377	906	621	95·4
1941	2,818	815	837	98·5
1942	3,104	758	1,006	99·4
1943	3,484	711	1,072	99·6
1944	3,953	715	1,082	99·6
1945	4,461	753	1,072	99·2
1946	4,846	847	1,251	97·6

From: London and Cambridge Economic Service *Bulletins* and Ministry of Labour *Gazette*.

Notes: Clearing banks—data for nine of present eleven banks throughout (excluding District and National); averages of twelve monthly figures (last making-up day each month since September, 1939, previously monthly averages). Employment—100 less unemployment percentage, taken as number of insurance books lodged in percentage of all insured workers in July. The series is broken in 1938 when two figures are given: (*a*) for insured workers 16—64, excluding agriculture and domestic workers, and total recorded unemployment amongst these (including "two months file"); (*b*) for insured workers 14—64, including agriculture and certain domestic workers, and registered unemployment amongst these (excluding "two months file"). Other changes in coverage have little effect on the percentages.

Object of table: To examine fluctuations in some indices of "business activity" over a period which includes boom, depression and war.

Some conclusions: Before 1939, bank advances and employment showed roughly comparable cyclical movements, but advances lagged behind, increasing only after employment had previously risen; bank investments showed a counter-cyclical variation, so that bank assets other than advances and investments were a fairly constant proportion of the total. These relations were disturbed by the war (*see* 3.3 and Figs. 2 and 4).

TABLE 7

Average Monthly Price of Eggs, England and Wales, 1929–38

s/d per 120

Month	1929	1930	1931	1932	1933	1934	1935	1936	1937	1938
Jan.	21/–	18/10	16/7	14/6	12/8	13/1	12/11	15/5	12/11	16/2
Feb.	20/–	18/–	14/1	12/3	13/11	11/5	11/7	14/2	13/8	14/6
March	17/–	11/8	10/11	9/4	9/1	8/–	8/5	9/7	10/10	10/4
April	11/8	11/2	9/4	8/8	7/7	7/11	7/8	8/6	9/–	10/4
May	12/11	10/11	9/3	8/4	8/–	7/8	8/6	9/6	9/8	11/4
June	12/11	11/9	9/6	9/6	9/6	9/3	9/11	10/6	11/9	12/2
July	16/2	14/9	12/2	11/10	10/10	10/1	11/9	12/8	14/10	15/4
Aug.	18/5	16/2	13/8	13/3	13/4	13/10	15/2	15/–	15/11	16/9
Sept.	19/7	17/–	15/1	15/6	14/8	13/–	15/–	15/6	17/8	18/5
Oct	26/3	22/8	18/9	17/10	16/7	16/9	17/3	20/5	19/10	19/11
Nov.	28/5	24/7	22/7	20/8	20/2	20/5	20/5	20/6	23/3	21/9
Dec.	25/10	20/1	16/5	16/2	17/5	17/1	19/4	18/10	21/3	17/8
Average	19/2	16/6	14/–	13/2	12/10	12/5	13/2	14/3	15/1	15/5

From: Agricultural Statistics.

Notes: Prices are at wholesale (wholesaler to retailer) and the quotation is English, ordinary pack, average of first and second qualities on town and country markets. Monthly figures are averages of weekly quotations (some months four weeks, other months five weeks) obtained by market reporters of Ministry of Agriculture.

Object of table: To show the variation in wholesale price of a commodity subject to severe seasonal fluctuations and, by isolation and elimination of the seasonal factor, to trace the general movement in price over the ten-year period of depression and recovery.

Some conclusions: There is a marked and regular seasonal movement in egg prices, which rise in summer and autumn and fall in winter and spring; the maximum price in November is about two and a half times the minimum of April. Apart from seasonal variations, the movement of prices in recession (1929–34) and recovery (1934–38) was a fairly regular cycle; by the summer of 1938 the general level of prices was still short of the 1929 record (*see* 8·4).

TABLE 8

Agricultural Rents and Crop Yields, S.E. England

County	Rents, 1941		Yields, 1929–38	
	Acreage (000 acres)	Rent (shillings per acre)	Wheat (cwts. per acre)	Barley (cwts. per acre)
	(1)	(2)	(3)	(4)
Bedford	238	26	17·7	15·7
Berks.	314	23	16·4	14·1
Bucks.	365	26	17·1	14·6
Cambridge	251	22	16·6	15·3
Isle of Ely	205	48	22·0	20·3
Essex	676	23	18·3	17·2
Herts.	279	22	16·8	16·3
Hunts.	192	24	16·8	14·3
Kent	617	30	20·2	19·0
Leicester	446	30	17·7	15·4
Middlesex	31	44	18·2	15·6
Norfolk	938	24	19·0	17·0
Northants.	484	25	16·9	15·9
S. of Peterboro.	41	26	17·4	14·6
Oxford	380	21	16·4	14·4
Rutland	88	20	18·1	15·4
Suffolk E.	428	20	19·3	17·0
Suffolk W.	284	20	18·0	16·2
Surrey	177	31	16·2	13·6
Sussex E.	300	27	17·2	14·9
Sussex W.	221	27	17·5	14·6
Warwick	459	28	16·7	15·1

From: National Farm Survey of England and Wales (1946), and *Agricultural Statistics* (1939).

Notes: Acreage is of crops and grass and for all holdings. Rent is per acre under crops and grass and computed from a sample of holdings of five acres and over. The proportion of holdings sampled increases with the size of holding from 5 per cent for holding of five and under twenty-five acres to 100 per cent for holdings of 700 acres and over. Yields are per acre under crop, averaged over ten years from annual estimates made by crop reporters of Ministry of Agriculture.

Object of table: To show the variation in rents and crop yields from one county to another, and to determine what degree of correlation exists between crop yields and rent.

Some conclusions: The counties in this region of England show large variations in all factors shown. There is a close relation, as expected, between yields of wheat and barley, but little correlation between either and rent. The factors influencing the level of rent must be numerous and complex (*see* 7.1 and Figs. 17 and 18).

TABLE 9

Distribution of Wholesale Prices, U.K.
(*Statist* Index Number of Wholesale Prices, Average 1867–77 = 100)

A. Price relatives, rounded to nearest 1 per cent of each of the forty-five commodity series in the *Statist* index (from contribution by Editor of *The Statist* in *Jour. Roy. Stat. Soc.*, 1946):

	In original order					In ascending order of magnitude				
1938	53	116	70	54	36	19	56	81	95	131
	70	98	176	139	38	24	57	81	98	133
	67	102	161	77	137	35	59	81	102	136
	93	133	56	93	81	36	60	83	105	137
	81	131	185	88	54	38	67	86	106	139
	86	92	83	60	109	53	70	88	109	161
	95	24	117	35	57	54	70	92	110	171
	106	19	171	81	79	54	77	93	116	176
	105	59	55	110	136	55	79	93	117	185
1945	113	142	116	160	97	29	108	129	160	218
	114	152	217	466	108	42	109	132	167	219
	87	153	218	219	236	70	113	134	176	229
	229	196	83	207	202	83	114	142	192	236
	176	192	286	153	86	83	116	142	196	272
	132	121	142	107	117	86	117	142	198	286
	118	42	198	70	109	87	118	152	202	325
	272	29	325	129	83	97	121	153	207	393
	134	127	142	167	393	107	127	153	217	466

B. First distribution of Price Relatives

Range of price relatives	No. of relatives	
	1938	1945
0–24	2	–
25–49	3	2
50–74	11	1
75–99	13	5
100–124	7	9
125–149	5	7
150–174	2	5
175–199	2	4
200–224	–	5
225–249	–	2
250–274	–	1
275–299	–	1
300–499	–	3
Total	45	45

C. Second distribution of Price Relatives

Range of price relatives	No. of relatives	
	1938	1945
Under 50	5	2
50–69	9	–
70–79	4	1
80–89	6	4
90–99	5	1
100–109	4	3
110–119	3	5
120–129	–	3
130–149	5	5
150–199	4	9
200–249	–	7
250–299	–	2
300 & over	–	3
Total	45	45

TABLE 10

Family Income and Expenditure on Food, Hamburg and Bremen, 1927–8
(RM per year)

A. Income and expenditure on food, rounded to the nearest RM, for each family in a sample of ninety families, arranged in ascending order of income:

Income	Food expend.	Income	Food expend.	Income	Food expend.	Income	Food expend.	Income	Food expend
2,311	1,034	2,910	1,179	3,248	1,530	3,502	1,740	3,844	1,435
2,395	1,003	2,910	1,533	3,259	1,475	3,516	1,261	3,860	1,615
2,395	1,120	2,936	1,390	3,270	1,480	3,535	1,579	3,916	1,570
2,427	1,146	2,971	1,293	3,270	1,521	3,537	1,322	3,957	1,886
2,440	1,120	2,982	1,643	3,276	1,462	3,541	1,835	4,018	1,811
2,516	1,300	2,984	1,319	3,281	1,477	3,549	1,199	4,037	1,701
2,536	1,020	2,989	1,405	3,298	1,746	3,560	1,446	4,056	1,528
2,636	1,159	3,005	1,279	3,329	1,389	3,563	1,442	4,168	1,722
2,683	1,288	3,012	1,184	3,351	1,289	3,609	1,652	4,216	1,902
2,744	1,375	3,057	1,552	3,351	1,464	3,616	1,446	4,226	1,796
2,752	1,422	3,058	1,286	3,351	1,669	3,699	2,078	4,241	1,285
2,808	1,241	3,099	1,413	3,357	1,318	3,707	1,634	4,287	1,854
2,820	956	3,108	1,298	3,382	1,773	3,717	1,696	4,295	2,268
2,825	1,224	3,134	834	3,384	1,396	3,726	1,746	4,323	1,961
2,845	1,239	3,166	1,378	3,403	1,830	3,764	1,436	4,379	2,168
2,846	1,493	3,167	1,368	3,411	1,407	3,790	1,408	4,407	1,726
2,866	1,466	3,198	1,661	3,467	1,827	3,801	1,648	4,512	1,544
2,900	1,322	3,236	1,105	3,492	977	3,815	1,842	4,828	2,018

From: Einzelschriften zur Statistik des Deutschen Reiche, Nr. 22 (1932).

Notes: Income and food expenditure in the year from March, 1927 to February, 1928 inclusive, as returned by families selected in a sample of working-class households. Families included are those which contain husband and wife, with or without children and other members at home, and which made returns for the whole year.

B. Income Distribution

Range of incomes	No. of families	"Normal" distribution[1]
2,300–2,499	5	4·7
2,500–2,699	4	4·8
2,700–2,899	8	7·5
2,900–3,099	13	10·3
3,100–3,299	13	12·4
3,300–3,499	11	13·1
3,500–3,699	11	12·1
3,700–3,899	9	9·8
3,900–4,099	5	7·0
4,100–4,299	6	4·2
4,300–4,499	3	2·3
4,500–4,699	1	1·1
4,700–4,899	1	0·7
Total	90	90 0

[1]*See* 9.7.

C. Cumulative Income Distribution

Income	No. of families
Under 2,500	5
,, 2,700	9
,, 2,900	17
,, 3,100	30
,, 3,300	43
,, 3,500	54
,, 3,700	65
,, 3,900	74
,, 4,100	79
,, 4,300	85
,, 4,500	88
,, 4,700	89
,, 4,900	90

D. Distribution of Income and Food Expenditure

Food expenditure	Income								
	2,300–2,599	2,600–2,899	2,900–3,199	3,200–3,499	3,500–3,799	3,800–4,099	4,100–4,399	4,400 & over	Total
	No. of families								
800–999	–	1	1	1	–	–	–	–	3
1,000–1,199	6	1	2	1	1	–	–	–	11
1,200–1,399	1	5	9	4	2	–	1	–	22
1,400–1,599	–	3	4	8	6	3	–	1	25
1,600–1,799	–	–	2	3	5	3	2	1	16
1,800–1,999	–	–	–	2	1	3	3	–	9
2,000 and over	–	–	–	–	1	–	2	1	4
Total	7	10	18	19	16	9	8	3	90

TABLE 11

Age Distribution and Mortality, England and Wales, 1938

Age (years)	Mid-year population				Deaths			
	Number (000's)		Percentages		Number		Per 1,000 of pop.	
	England and Wales	S.W. England	England and Wales	S.W. England	England and Wales	S.W. England	England and Wales	S.W. England
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
0–	2,818	131·6	6·8	6·3	42,940	1,699	15·2	12·9
5–	6,025	278·0	14·6	13·3	9,050	338	1·5	1·2
15–	6,572	302·8	15·9	14·5	14,300	632	2·2	2·1
25–	6,817	322·1	16·5	15·5	18,310	782	2·7	2·4
35–	6,034	300·3	14·6	14·4	24,150	1,117	4·0	3·7
45–	5,134	263·5	12·5	12·7	43,290	2,099	8·4	8·0
55–	4,240	237·5	10·3	11·4	79,960	4,099	18·9	17·3
65–	2,561	168·5	6·2	8·1	117,780	6,848	46·0	40·6
75	1,014	78·3	2·5	3·8	129,050	9,316	127·3	119·0
Total	41,215	2,082·6	100·0	100·0	478,830	26,930	11·6	12·9

From: Registrar-General's *Statistical Review* (1938, Part I).

Notes: Mid-year population (resident) as estimated by Registrar-General. Deaths as registered in year including non-civilians. S.W. England comprises the counties of Cornwall, Devon, Dorset, Somerset and Wilts.

Object of table: To compare the mortality rates in one area (S.W. England) with those in the country as a whole.

Some conclusions: The "crude" death rate in S.W. England is greater than that in the whole country; but the rates in each individual age group are lower in S.W. England. Hence mortality is lower in this area than in the country as a whole and the higher "crude" death rate arises because of the older population of the area. The "crude" death rate needs to be corrected to eliminate the effect of the age distribution of the population (*see* 6.8).

TABLE 12
Distribution of Private Incomes, Before Tax, U.K.

Incomes which can be allocated to income ranges	1938		1945	
	No. of incomes (*a*) 000's	Total income £ millions	No. of incomes (*a*) 000's	Total income £ millions
£	(1)	(2)	(3)	(4)
250–	1,745	595	5,400	1,895
500–	500	350	1,650	1,140
1,000–	195	270	410	535
2,000–	97	360	124	462
10,000 and over	8	170	8	138
Total allocated (£250 and over)	2,545	1,745	7,592	4,170
Incomes under £250 (*b*)	...	2,681	...	3,565
Unallocated private income	...	681	...	1,480
Total private income		5,107		9,215

From: National Income and Expenditure of the U.K. 1938 *to* 1946 (Cmd. 7099, 1947).

Notes: Estimates made in the Central Statistical Office, using Inland Revenue and other official data. Total private income is the sum of rent, interest, profits, wages and salaries received by persons, together with pay and allowances of Forces, undistributed profits (including tax liabilities) and transfer payments such as pensions, unemployment and sickness benefits.
(*a*) Married couples counted as individuals.
(*b*) Including all transfer payments except war gratuities.
...Not available.

Object of table: To trace changes in the personal distribution of incomes (before payment of tax) arising from the war of 1939–45.

Some conclusions: Of all allocated private income, 60 per cent went to individuals each earning less than £250 in 1938, but only 45 per cent in 1945. The numbers of individuals receiving incomes of £250 and over increased threefold between the two years, the largest increases being in the numbers receiving moderate incomes of £250 to £1,000. The main, but not the complete, explanation is the general rise in prices and incomes from 1938 to 1945.

TABLE 13

Coal Mines Classified by Size and by Output per Manshift, Great Britain, 1945

No. of wage-earners employed	Output per manshift (cwts.)									
	Under 10	10–	15–	20–	25–	30–	35–	40–	45 and over	Total
	Number of mines									
1–19	86	87	99	44	40	14	6	5	12	393
20–49	8	31	55	48	15	13	2	2	–	174
50–99	3	28	28	26	18	4	1	1	–	109
100–249	9	24	64	29	14	7	–	–	–	147
250–499	10	60	76	47	14	7	–	–	–	214
500–749	5	37	71	50	19	7	–	–	–	189
750–999	1	15	60	26	10	2	1	–	–	115
1,000–1,499	2	14	42	47	12	8	2	1	1	129
1,500–1,999	–	8	21	20	11	2	1	–	–	63
2,000–2,499	–	1	5	8	4	1	–	–	–	19
2,500–2,999	–	–	3	7	4	–	–	–	–	14
3,000 and over	–	–	2	1	1	–	–	–	–	4
Total	124	305	526	353	162	65	13	9	13	1,570

From: Ministry of Fuel and Power *Statistical Digest*, 1945 (Cmd. 6920, 1946).

Notes: Size of mine measured by total number of wage-earners employed, underground and on surface, at 15th December, 1945 (or nearest working day). Output per manshift by division of annual output of coal from the mine by number of manshifts worked in year, latter being number of wage-earners in each shift summed over all shifts worked. Calculation on over-all basis, including both underground and surface workers. Allowance made for manshifts not attributable to coal-mining in mines producing significant quantities of minerals other than coal.

Object of table: To examine the relation of size of mine to productivity, as measured by output per manshift in the mine.

Some conclusions: In a comparison of groups of mines with various levels of productivity, the representative (average) size of mine first increases with increasing productivity but then decreases in the groups of mines with highest productivity (25 cwts. and more per manshift); the relation is, therefore, not one of steady progression. The largest mines have middling productivity while very low and very high productivity are found in small mines (*see* 3.7 and 7.2).

TABLE 14

FAMILY SIZE AND ACCOMMODATION, KENSINGTON, 1929–30

No. of rooms	Number of persons								Total	Mean rent (shillings)
	1	2	3	4	5	6	7	8 and over		
	Number of families									
1	41	22	7	8	–	–	–	–	78	5·7
2	21	50	39	26	18	12	6	3	70	9·1
3	6	32	34	25	12	12	7	3	131	12·0
4	2	7	18	19	8	10	3	3	70	14·2
5	–	–	5	5	5	5	1	3	24	18·5
6 and over	–	1	1	1	–	2	1	–	6	22·8
Total	70	112	104	84	43	41	18	12	484	10·5

From: New Survey of London Life and Labour, Vol. VI (1934).

Notes: Sample of households taken in Kensington, 1929–1930, comprising 484 working-class families, as shown here, together with 517 middle-class families and sixty-two families of unknown status. Rooms occupied defined as in the population census, including kitchens only if used for living. Rent taken as net rent after deduction of receipts from sub-letting.

Object of table: To examine the relation of family size and rent paid to the size of dwelling occupied.

Some conclusions: Though the number of rooms occupied is generally greater for larger families, there is considerable variation in the accommodation of families of any given size; mean rent paid increases progressively, and significantly, as the size of dwelling increases (*see* 7.7 and 9.5).

APPENDIX II

A SHORT READING LIST

As general introductions to the use of statistics in economic and social problems:

1. M. A. Abrams, *The Condition of the British People*, 1911–1945 (Fabian Society, 1946);
2. A. M. Carr-Saunders and D. C. Jones, *The Social Structure of England and Wales* (Oxford, 2nd. Ed., 1937);
3. L. H. C. Tippett, *Statistics* (Home University Library, 1943);

and as elementary accounts of statistical methods:

4. L. R. Connor, *Statistics in Theory and Practice* (Pitman, 3rd Ed., 1938);
5. E. C. Rhodes, *Elementary Statistical Methods* (Routledge, 1933).

As general textbooks on statistical methods and applications, mainly non-mathematical:

6. F. E. Croxton and D. J. Cowden, *Applied General Statistics* (Prentice-Hall, N.Y., and Pitman, 1947);
7. J. G. Smith and A. J. Duncan, *Elementary Statistics and Applications* (McGraw-Hill, N.Y., 1944);

and as textbooks of more mathematical content:

8. A. L. Bowley, *Elements of Statistics* (Staples, 6th Ed., 1937);
9. G. U. Yule and M. G. Kendall, *An Introduction to the Theory of Statistics* (Griffin, 13th Ed., 1945).

As introductions to the theory of statistics from the mathematical angle:

10. A. C. Aitken, *Statistical Mathematics* (Oliver and Boyd, 4th Ed., 1945);

11. P. G. Hoel, *Introduction to Mathematical Statistics* (John Wiley, N.Y., and Chapman and Hall, 1947);
12. C. E. Weatherburn, *A First Course in Mathematical Statistics* (Cambridge, 1946);

and as texts designed primarily for the experimentalist but also of more general interest:

13. R. A. Fisher, *Statistical Methods for Research Worker* (Oliver and Boyd, 10th Ed., 1946);
14. C. H. Goulden, *Methods of Statistical Analysis* (Wiley, N.Y., 1939);
15. E. F. Lindquist, *Statistical Analysis in Educational Research* (Houghton Mifflin, Boston, 1940);
16. G. W. Snedecor, *Statistical Methods* (Iowa State College Press, 4th Ed., 1946);
17. L. H. C. Tippett, *The Methods of Statistics* (Williams and Norgate, 3rd Ed., 1941).

As accounts of special sections of statistical theory:

18. H. T. Davis, *The Analysis of Economic Time Series* (Bloomington, Indiana, 1941);
19. J. G. Smith and A. J. Duncan, *Sampling Statistics and Applications* (McGraw-Hill, N.Y., 1945);
20. A. A. Tschuprow, *Principles of the Mathematical Theory of Correlation* (Hodge, English Ed., 1939);
21. H. Wold, *A Study in the Analysis of Stationary Time Series* (Uppsala, 1938);
22. F. Yates, "A Review of Recent Statistical Developments in Sampling and Sampling Surveys" (*Jour. Roy. Stat. Soc.*, 1946).

Finally, as advanced textbooks on mathematical statistics:

23. H. Cramér, *Mathematical Methods of Statistics* (Princeton, 1946);
24. M. G. Kendall, *The Advanced Theory of Statistics* (Griffin, Vol. I, 2nd Ed., 1945; Vol. II, 1946);
25. S. S. Wilks, *Mathematical Statistics* (Princeton, 1944).

INDEX

STATISTICAL METHODS

G

H

I

L

M

N

O

P

Q

R

S

T

V

W

APPLICATIONS

AUTHORS AND SOURCES

A

Abrams, M. A., 206
Agriculture, Ministry of, *Agricultural Statistics*, 35, 36, 117, 190, 192
—*Journal*, 117
—*National Farm Survey*, 11, 192
Aitken, A. C., 206
Allen, R. G. D. and Bowley, A. L., 39

B

Bank of England, *Return*, 30
—*Statistical Summary*, 31
Barna, T., 30, 34
Beveridge, W. H. (Lord), 37
Booth, Charles, 10, 40
Bowley, A. L., 10, 29, 34, 110, 111, 137, 186, 206
Bowley, A. L. and Smith, Katie C., 146

C

Campion, H., 30
Carr-Saunders, A. M., 33
—and Jones, D. C., 206
Central Statistical Office, *Annual Abstract of Statistics*, 27
—*Monthly Digest of Statistics*, 27, 31, 34, 35, 36, 38, 117, 180
—*National Income and Expenditure*, 28, 29, 30, 38, 114, 118, 178, 200

D

E

F

G

H

I

K

L